DOCUMENT RESUME

ED 090 696 EA 006 148

AUTHOR	Gage, N. L., Ed.
TITLE	Mandated Evaluation of Educators: A Conference on California's Stull Act. (Stanford, California, October 12-14, 1972.)
INSTITUTION	Stanford Univ., Calif. Stanford Center for Research and Development in Teaching.
PUB DATE	73
NOTE	366p.
AVAILABLE FROM	Education Resources Division, Capitol Publications, Inc., Suite G-12, 2430 Pennsylvania Avenue, N. W., Washington, D. C. 20037 ($9.00)
EDRS PRICE	MF-$0.75 HC-$17.40 PLUS POSTAGE
DESCRIPTORS	*Administrative Personnel; *Administrator Evaluation; Conference Reports; Educational Objectives; Evaluation Methods; *Legal Problems; *Measurement Techniques; *State Legislation; Statistical Analysis; *Teacher Evaluation; Teachers; Test Reliability; Test Validity
IDENTIFIERS	*California; Stull Act

ABSTRACT

The Stull Act (after its principal sponsor, Assemblyman John Stull) requires all school districts in California to adopt a uniform set of guidelines for use in evaluating the professional competence of their certificated personnel. The act went into effect on September 1, 1972. In the spring of 1972, it became apparent from newspaper articles that the school districts of California were seriously concerned and perhaps panicky about the Act. In recognition of the complex and sensitive problems the Act entailed -- problems in measurement, evaluation, and statistical analysis; in the definition of achievement; and in law and administration, -- the Stanford Center for Research and Development in Teaching organized a conference of experts in the relevant disciplines to discuss the Act. This book contains the proceedings of that conference. (Author/JF)

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED 090696

Mandated Evaluation Of Educators:

A Conference on California's Stull Act

Edited by N. L. Gage
Stanford Center for Research and
Development in Teaching

EA 006 148

ERD Education Resources Division Capitol Publications, Inc.
Suite G-12, 2430 Pennsylvania Avenue, N.W., Washington, D.C. 20037

ED 090696

MANDATED EVALUATION OF EDUCATORS:
A CONFERENCE ON CALIFORNIA'S STULL ACT

Edited by

N. L. Gage

Stanford Center for Research and Development in Teaching
School of Education, Stanford University
Stanford, California
1973

PREFACE

The Stull Act (after its principal sponsor, Assemblyman John Stull) requires all school districts in California to adopt a uniform set of guidelines for use in evaluating the professional competence of their certificated personnel. Because the Act requires the evaluation of teachers, administrators, and other kinds of professional staff, it formally imposes on all California school districts a task that has previously been performed only informally.

The problem of evaluating certificated personnel in education is an old one. Research workers and pioneering school districts have long been making forays into this difficult area. The Stull Act requires all school districts to tackle the problem, to do so quickly, and to take the evaluations seriously.

The Act--Assembly Bill 293, passed in the 1971 session of the California legislature--went into effect on September 1, 1972. In the spring of 1972, it became apparent from newspaper articles that the school districts of California were seriously concerned--perhaps "panicky" is not too strong a word--about the Act. Practical school people saw all sorts of difficulties, and their concern was shared by research workers and technical specialists in evaluation.

During this period, the Program on Teaching Effectiveness of the Stanford Center for Research and Development in Teaching was planning its work for the next few years. That program was aimed at the development of a model teacher training system. One component of that system was to be a set of techniques for evaluating teachers--their behaviors, characteristics, and effectiveness. Hence the SCRDT program became specifically concerned with the Stull Act, in recognition of the complex and sensitive problems it entailed--problems in measurement, evaluation, and statistical analysis, in the definition of achievement, and in law and administration.

The SCRDT plan therefore included the organization of a conference of experts in the relevant disciplines and practices from the State's Department of Education, school districts, organizations of educators, and universities. The papers and discussions of this conference were to be made available to the school districts of the State and other interested parties. These proceedings were expected to have implications extending well beyond California.

The conference was held at Stanford University on October 12-14, 1972. The scope and topics of the papers to be presented were planned in May 1972 by an ad hoc committee consisting of Alan C. Hopewell, Association of California School Administrators; James A. Klawitter, California School Boards Association; Garford G. Gordon, California Teachers Association; Marvin C. Alkin, Center for the Study of Evaluation, UCLA; Gerald Pangburn, then of the Department of Education, State of California; David C. Berliner, Far West Laboratory for Educational Research and Development; and N. L. Gage, Chairman, Stanford Center for Research and Development in Teaching. The Committee was joined in its work by additional representatives from some of these organizations, namely, Frank Brunetti, Robert N. Bush, and Richard J. Shavelson of SCRDT; J. W. Stallings of the University of California; and Robert Stahl of the California Teachers Association. Advice and suggestions as to the selection of authors of the invited papers were received from all these persons and many others.

The conference sessions were capably chaired by Robert K. Rice, Department of Education, State of California; Robert Stahl, California Teachers Association; Arthur North, Santa Barbara School Districts; Barbara Hatton, Stanford University; Robert Ehlers, Department of Education, State of California;

Gerald Pangburn, Sierra Joint Union High School District; Frank Bettis, Rio Linda High School; and Walter T. Coultas, Department of Education, State of California.

Approximately 300 educators attended the conference. Their informal discussions (not presented here) were lively and cogent. The proceedings were recorded and transcribed. The papers and the transcripts of the formal discussions were edited and then reviewed by their authors. These materials, together with opening remarks and a copy of the Stull Act, make up the present volume.

In making arrangements for the conference, I was helped greatly by Margaret Bander and Ann Gladstone, SCRDT secretaries, and by Christopher Clark, program research assistant. The primary editing of the papers and discussions was capably performed by Frances H. Maurier. Bruce Harlow, Betty K. Smith, and Sumi Kawasaki of the SCRDT Publications and Dissemination Unit gave their services unstintingly in arranging for the editing, manuscript preparation, and publication of the present volume.

Although each reader must, of course, judge for himself, the proceedings as here presented seem to me to comprise a searching and constructively critical analysis of the Stull Act. These papers and discussions should help materially in complying and dealing in other ways with the Act. Since almost all of the points made here apply outside California, these proceedings should also be of interest to readers in other states where similar legislation is being considered or adopted.

N. L. Gage, Director
Program on Teaching Effectiveness
Stanford Center for Research and
Development in Teaching

CONTENTS

List of Speakers ix

Opening Remarks 1

Arthur P. Coladarci
Robert N. Bush

Legal Problems in Implementing the Stull Act 5

Thomas A. Shannon

Discussants: Jack Friedenthal 32
Peter T. Galiano 38
Raoul Teilhet 42

Statistical and Measurement Problems in Implementing the Stull Act 53

Gene V Glass

Discussants: Henry M. Levin 99
Henry F. Kaiser 104
W. James Popham 107

Problems of Defining Educational Objectives and Standards in Implementing the Stull Act 111

W. James Popham

Discussants: Elliot W. Eisner 133
John K. Hemphill 140

Evaluating Nonteaching Certificated Personnel 144

Max G. Abbott

Discussants: Lewis Allbee 160
James A. Klawitter 161
Duncan Sprague 166

The Assessment of Proper Control and Suitable Learning Environment 169

Ned A. Flanders and Greta Morine

Remarks: Greta Morine 234
Discussants: John Young 236
Keith Waite 240

The Use of Evaluation to Improve Performance 244

Marvin C. Alkin and Eva L. Baker

Discussants: Thomas W. Butterworth 283
Bernard H. McKenna 289

Evaluating Evaluation Systems Designed to Implement the Stull Act 296

Garford G. Gordon

Discussants: James F. Garvey 325
Robert L. Forbes 327

Obeying and Improving the Mandate to Evaluate Educators: Some Lessons from the Conference on the Stull Act 334

N. L. Gage

Appendix: Assembly Bill No. 293 (The Stull Act) . . . 349

LIST OF SPEAKERS

MAX G. ABBOTT, Director, Center for the Advanced Study of Educational Administration, University of Oregon

MARVIN C. ALKIN, Director, Center for the Study of Evaluation, University of California, Los Angeles

LEWIS ALLBEE, Superintendent, Barstow (Calif.) Unified School District

EVA L. BAKER, Professor of Education, University of California, Los Angeles

ROBERT N. BUSH, Director, Stanford Center for Research and Development in Teaching, Stanford University

THOMAS W. BUTTERWORTH, Consultant in Research and Pupil Personnel, Los Angeles County Schools Office

ARTHUR P. COLADARCI, Dean, School of Education, Stanford University

ELLIOT W. EISNER, Professor of Education and Art, Stanford University

NED A. FLANDERS, Director, Program on Effective Teacher Education, Far West Laboratory for Educational Research and Development, San Francisco, California

ROBERT L. FORBES, Principal, Canyon High School, and Coordinator of Secondary Education, Castro Valley (Calif.) Unified School District

JACK FRIEDENTHAL, Professor of Law, Stanford University

N. L. GAGE, Director, Program on Teaching Effectiveness, Stanford Center for Research and Development in Teaching, Stanford University

PETER T. GALIANO, General Counsel, California Teachers Association, Burlingame, California

JAMES F. GARVEY, District Superintendent, Glendora (Calif.) Unified School District

GENE V GLASS, Co-Director, Laboratory for Educational Research, University of Colorado

GARFORD G. GORDON, Research Executive, California Teachers Association, Burlingame, California

JOHN K. HEMPHILL, Executive Director, Far West Laboratory for Educational Research and Development, San Francisco, California

HENRY F. KAISER, Professor of Education, University of California, Berkeley

JAMES A. KLAWITTER, Field Executive, California School Boards Association, Sacramento, California

HENRY M. LEVIN, Associate Professor of Education, Stanford University

BERNARD H. McKENNA, Professional Associate, Division of Instruction and Professional Development, National Education Association, Washington, D. C.

GRETA MORINE, Associate Professor of Education, California State University, Hayward

W. JAMES POPHAM, Professor of Education, University of California, Los Angeles

THOMAS A. SHANNON, Schools Attorney, San Diego City Schools and Community Colleges; Legal Counsel, Association of California School Administrators

DUNCAN SPRAGUE, Instructional Services Consultant, California Teachers Association, Stockton, California

RAOUL TEILHET, President, California Federation of Teachers, AFL-CIO, Burbank, California

KEITH WAITE, Superintendent, Tulare (Calif.) Union High School District

JOHN YOUNG, Executive Director, Mt. Diablo Unified District Education Association, Walnut Creek, California

INTRODUCTORY REMARKS

ARTHUR P. COLADARCI
Stanford University

I welcome you on behalf of the School of Education and I thank all those who made the conference possible. The immediate and possible future consequences of the Stull Act are very real and I have some fear that we may respond to this legislation much as we have to other kinds of pressing legislation and needs--that is, by being very compulsive about reliability and clarity but largely ignoring the questions of criteria and validity. I like the heuristic definition of the problem of teaching and its evaluation as comprising an infinite number of transactions--between pupils and teachers, and so on. I have some fear that, if we are not careful, we may once more focus on meretriciously simple resolutions of the criterion problem and may ignore entirely what the schools are really all about. The danger here lies in focusing on just one transaction, that between a pupil and something called a California Achievement Test or a Stanford Achievement Test battery.

Thus, the subject in which you are engaged here is critical not only to California but, I suspect, to the whole country. If the problems discussed at this conference can be adequately resolved, this meeting may well prove a landmark in education.

ROBERT N. BUSH
Stanford University

Ladies and gentlemen, we are here on serious business. Last night I read the Saturday Review's section for this month on education. One of the leading articles is headlined: "My kid will get her high-school diploma this year. She can't read the want ads. She can't fill out a job application, and if she gets a job she'll probably be fired because she can't write legibly or make change for a dollar. I've got a good mind to sue somebody." The article goes on to state that public education is the nation's largest consumer industry, and it looks as if the schools could soon become a target for militant consumer movements. In cases where the quality of education is demonstrably poor, there is reason to believe that consumers may legitimately take action in the courts. Students, parents, taxpayers, and for that matter Ralph Nader may well claim that the principles of law that govern business, industry, and some professions also extend to education. The article then proposes a hypothetical case of suit against the schools. On the opposite page is a one-page advertisement for books totaling $42.80 (which, if you join now, you can get for $1 each) on the subjects of accountability, deschooling society, and so on.

I think it is safe to say that there is a significant groundswell concerning accountability in education in this country. Dr. Levin (whom you will hear as a discussant a little later) has written a most perceptive paper on the subject of accountability which I hope you will have an opportunity to read. He points out a number of complications in the problem of accountability. He summarizes by saying we ought to be thinking of accountability to whom and for what, and that these questions have not been clarified. I would add two other statements. Who is accountable? And in addition to for what and to whom, also how? It seems to me that therein is one of the great problems--who is to be accountable, and in what manner? What would be the reaction, I ask you, if the California Legislature passed an act calling for evaluation of all medical personnel with regard to how well they cured cancer, the common cold, heart disease, or diabetes? Would we fire the doctors if they couldn't pass the evaluation tests?

I think the serious question facing us is this: What is within our competence with regard to being accountable? The Stull Act is on the books; it is law, and that is a sobering reality. There has been a great deal of discussion, but what we need, it seems to me, is much more light and less heat, much more technical competence and less rhetoric, as we discuss the problem. It was indeed out of this background that we came to feel this conference ought to be called, to bring to bear the technical competence that is available in attempting to carry this legislation out.

As Dr. Gage has suggested, the eyes of the nation are upon us because there are at least a dozen other states in which similar legislation is pending, some of it with the exact wording of the California law. We have a great people in California, and we have great technical competence to bring to bear on the problem. In addition to our traditional technical competence, during the last few years the federal government has created a network of regional laboratories and research and development centers which are designed to study key educational policy questions. There are four such centers and laboratories in California, one in higher education and the other three dealing mostly with the work of the public schools--one at UCLA, one soon to be moving to San Francisco, and one here at Stanford. These three groups are pleased to join with teachers and administrators--the practical people working in the field--to try to cope with this important problem. Whatever the consensus--or lack thereof--that arises out of this meeting, it should be important for the future of the Stull Act and for the improvement of education in California.

LEGAL PROBLEMS IN IMPLEMENTING THE STULL ACT

THOMAS A. SHANNON
San Diego City Schools and Community Colleges and
Association of California School Administrators

The 1971 California State Legislature enacted two comprehensive new laws governing the evaluation of public school teaching and administrative personnel. One of these laws, originally known as the "Stull Bill" after its principal author, Assemblyman John Stull, governs the evaluation of teachers and administrators employed in kindergarten through grade 12;* the other new law, known as the "Rodda Bill" because it was authored by Senator Albert Rodda, applies to the evaluation of teachers and administrators employed in the community colleges of California, which we all recognize is the new name for the "junior colleges" in which students enroll for grades 13 and 14.*

---

*The Stull Act has been described as a certificated employee "tenure law" and as a certificated employee "evaluation law." Both descriptions are accurate. The Act is a certificated employee tenure law in that it prescribes the legal grounds upon which a public school certificated employee in California may be dismissed and establishes the procedures which must be used to determine if such legal grounds for dismissal actually exist. That is all "tenure" is. Stated in a simple mathematical formula:

Teacher tenure = Specific, known reasons which are the exclusive grounds on which a teacher's employment may be terminated. + Specific, known procedure for holding a hearing to determine whether there are valid reasons to terminate a teacher's employment.

The Act is a certificated employee evaluation law in that it sets forth specific requirements for a certificated employee evaluation procedure in local public school districts in California. Stated another way, the Act overhauls the old certificated employee dismissal procedures and introduces a new system of certificated employee evaluation in California.

Appendix A to this paper is a "bird's-eye view" of the Stull Act affecting the evaluation of teachers and administrators in the public elementary and secondary schools of California. It is this Appendix to which we will refer during our discussion of the legal aspects of the evaluation of public school certificated employees under the Stull Act.

Someday, the 1970's may become known as the "decade of accountability" in public education. The frustration over the growing effectiveness of teacher collective negotiations; the burgeoning tax rate to support federal, state and local governments, including school districts; the dramatic introduction of computer sciences to many important areas of private business and governmental operation; the widespread dissatisfaction of minority race groups with the public school systems; the lack of general understanding of our citizens of the encompassing and pervasive effect of the "children's revolution" which has seen courts and legislatures bestow rights and "degrees of status" on young people absolutely unheard of in years past; the cancerous erosion of family organization and strength; the affluence of children today which has resulted in a substantial part of our private business economy catering to the whims of youngsters; the disillusionment of young people with the "Protestant ethic" of "work" and "storing up for the future" (which we believe made our nation great but which many young people claim has made us slaves to materialism; when we vigorously say that a "man is what he does," they argue with equal vigor that "being" is more important than "doing"); and the attempt to loosen the so-called arbitrary restraints upon the conduct of man which for generations have been known as the "moral code" in favor of a permissive philosophy. . .all of these have merged and focused on the public schools because that is where the children are, and this has resulted in a demand that school teachers must do a "better job" with the children. This demand is heard loudly and clearly by state legislatures, which are really the master school boards in their states. As a result, in California and other states of the nation, the state legislatures have zeroed in on teacher evaluation as the means of bringing more accountability into the schools.

---

*Legislative Proposal AB 1192 was approved by the 1972 Regular Session of the California Legislature but vetoed by the Governor. Generally, the veto was based on the Governor's view that local school boards should retain the authority to determine how test scores should be used in the certificated employee evaluation process. In summary, AB 1192 would have:

a. Established, as the official name of the Stull Act, the title "Stull-Rodda Professional Competence Act."

b. Changed the concept of "uniformity." Instead of requirint that the written evaluation guidelines must be "uniform" in the school district, it would have specifically provided that such guidelines for evaluation must be "uniform" in the school district or in a territory within the school district or in a school itself. The net effect of this amendment is to permit considerably more flexibility than many thought was permitted under the Stull Act as it was originally enacted.

c. Prohibited the use of the norms identified by test publishers as criteria for the evaluation and assessment of certificated employees under the Stull Act.

Other 1972 bills which would have amended the Stull Act but apparently died during the session were SB 372, SB 379, and SB 1055.

At first blush, evaluation of teachers and administrators appears to be strictly a personnel management problem. But when it is realized that the ultimate enforcement of the evaluation process involves dismissing or demoting deficient employees, the legal implications of evaluation become crystal clear. That is, when a school board, acting as a governmental employer, undertakes to assess the competency of a person to practice his profession in accord with his state-granted license and, as a result of such assessment, to discipline or dismiss that person, or to reward one person and not another in the same class of employees, the school board indeed treads into legal thickets and brambles. In a very real sense, school boards have two burdens to carry. (1) Because a school board is a local governmental agency, the courts do not hesitate to examine carefully the reasonableness of any of board actions, including those which relate primarily to the school board's function as a substantial employer in the community. This is especially true if the dispute has constitutional dimensions, such as the First Amendment "free speech" rights of a teacher or school administrator. In short, a school board is never allowed to vent its spleen upon one of its employees, as may often happen with private employers dealing with their professional or management employees. (2) As a governmental employer, a school board must operate under an elaborate construct of laws which on the one hand constitutes an elaborate refinement of civil service job security and on the other hand represents in later years a deep thrust into collective negotiations, on the model of the private sector's labor-management experience. Today, school boards have all of yesterday's world of civil service protection and a great amount of today's world of organized labor in the private sector of our economy.

It is within this framework of the changing status of our youth, coupled with the changing dynamics of employer-employee relations in school districts, that the legal dimensions of certificated employee evaluation stand out in bold relief.

We shall begin our analysis with the law itself. The

Stull Act was enacted by the 1971 regular session of the California Legislature. During the current 1972 legislative session, four laws were enacted which affect the Act. In summary, these new laws were:

1. Chapter 10, Statutes 1972.

This new law changed the effective date of the evaluation portion of the Stull Act from March 4 to September 1, 1972. The change was made because of the problems of gearing up for Stull Act evaluations so quickly and the impact of such new evaluations upon probationary teacher nonreemployment cases arising just prior to March 15, 1972. This new law, or urgency measure, is reflected in Appendix A.

2. Chapter 535, Statues 1972.

This law exempts certificated personnel employed on an hourly basis in adult education classes from the provisions of the Stull Act requiring evaluation of certificated personnel. It is presently in effect.

3. Chapter 1013, Statutes 1972.

This law affords the parties in teacher dismissal hearings under the Stull Act the rights and duties of discovery of evidence enjoyed by litigants in a civil action brought in a Superior Court, except that all such discovery must be completed prior to one week before the date set for the hearing. This law becomes effective 91 days after adjournment of the 1972 regular session of the California Legislature.

4. Chapter 795, Statutes 1972.

This law will also have an impact upon implementation of the Stull Act. It is Assemblyman Ryan's "Master Teacher Act," which was an "urgency statute" and is now in effect. This law contains two parts. The first part, applicable only to school districts with a teaching staff of not less than 100 nor more than 650 and an average daily attendance (ADA) of not less than 2,500 nor more than 10,000 provides for a "master teacher selection panel," comprised of two teachers from another school district, to select "master teachers." A teacher so selected would receive $4,000 more in salary in recognition of his or her special competence. For the pilot program, state money is available. The other part of the new law establishes tenure rights of certificated employees who are not qualified to serve in a classroom. This group includes school psychologists, psychometrists, counselors holding only pupil personnel services credentials, etc. These types of employees gain tenure in these positions and, if promoted to an administrative position and later demoted, they have a right to revert back to that position. If the employee also may render service in the classroom, his tenure is that of a classroom teacher--otherwise, it is that of the special field in which he is qualified to serve.

These two aspects of the "Master Teacher" law impinge upon the Stull Act in that a new form of evaluation will be developed independently of the school board and administered by teachers who are entirely outside the jurisdiction of the school board or the school district's administration. Where the guidelines developed to implement the "Master Teacher Act" are different from the school board's guidelines enacted pursuant to the Stull Act, considerable confusion will arise and the differences will be interposed as a defense by certificated employees charged with incompetency for failure to meet the standards set by the school board. However, the granting of tenure rights to nonclassroom, certificated personnel seems to be consistent with the concept of the Stull Act that all certificated personnel, not only teachers, are subject to its evaluation procedures. Whether or not these new tenure rights comport with the direction in which the "Winton Act" seems to be going is quite another matter.

With this background in mind, and with the concept firmly fixed in our minds that the ultimate enforcement of evaluation depends on the legal nature of the evaluation guidelines adopted under the Stull Act, several specific legal questions come to mind. The analysis of these questions necessarily involves the ultimate enforcement concept. I mention this at the outset to avoid the impression of negativism; that is, there is a whole host of positive reasons why good evaluation of certificated personnel should be undertaken, and the description of these reasons is properly the province of school board members, administrators, and personnel management experts. It is the unhappy lot of school district legal counsel to advise that if the evaluation is not sound, a school board will never be able to enforce its standards through dismissal or other discipline of those who cannot meet the school district's criteria.

The several fundamental legal issues in implementing the Stull Act begin with the question:

What sanctions can be imposed against a school district for not implementing the Act?

To begin with, that district may as well forget forevermore any attempt to dismiss or otherwise discipline a certificated employee, administrator, or teacher for lack of competence. The Stull Act enjoins the school district to set standards; where there are no standards there can be no failure to live up to the standards. Viewed in another light, each school board member and administrator takes an oath of office to uphold the law of California; for them, no better reason should be needed to comply with the Act. Looked at from another perspective, a golden opportunity to devise standards for the most important endeavor in our society will be lost. Of course, any citizen or group of citizens or employees could obtain a writ of mandate from the local

Superior Court to compel the school board to establish evaluation procedures as required by the Stull Act. But the problem most likely will not be a refusal to implement the Act--rather, it will center on an inept or negligent implementation of the Act. And this brings us to the second question.

What degree of preciseness must characterize the school board's written guidelines adopted under the Stull Act?

The guidelines must serve as the exclusive legal basis for determining the level of professional competency of a teacher or administrator proposed for dismissal on grounds of incompetence. Accordingly, the guidelines must have been adopted by the school board in a public meeting; they must be clear in their intent and understandable in the language used; they must establish reasonable standards under all the circumstances; they must be applied in a consistent and uniform manner to the same class of employees; they must not contain any internal inconsistencies within the standards themselves or external inconsistencies when compared with other school board regulations, district administrative directives, regulations of the State Board of Education, or other laws of the Legislature or provisions of the federal or state constitutions; they must conform to all requirements specified in the Stull Act; and the school district must have made the regulations conveniently available to the person being evaluated. In short, the certificated employee involved must have had a reasonable opportunity to know and understand the standards being applied to him. In effect, the school board regulations implementing the Stull Act must have been drafted with virtually the same attention to clearly explaining the elements of expected conduct as is given to a penal statute in the Criminal Code. Because such important rights are at stake for the individual who is resisting being branded with the mark of incompetency in his chosen profession, the courts may be expected to resolve every ambiguity, every vague or incomplete sentence, every impossible requirement, and every undefined term against the school board in a dismissal or discipline action. In this kind of a legal skirmish, regardless of whether it is played out before a "professional competency panel" or a court, the entire organized administrative or teaching profession will unite behind an employee whose professional life is threatened because he did not abide by fuzzy or imprecise standards or procedures. And this takes us to the third question.

What is the relationship between the Stull Act and the Winton Act?

The Winton Act* governs employer-employee relations in the public school districts of California. Under this Act, school boards must meet and confer, upon request, with certificated employee organizations (or with a "certificated employee organization") on employer-employee relations matters, except so-called "educational policy" matters.

Under the Stull Act a school board is expressly required to avail itself of the advice of certificated instructional personnel in the development of evaluation guidelines. The issue becomes this. Is the school board required, in the initial development and continuing improvement of evaluation guidelines, to deal exclusively with the Winton Act certificated employee organization representative? I am of the view that the answer to this question must be "no." This conclusion is based primarily on the facts that: (1) the Stull Act makes no mention of the

---

*Education Code Section 13080 et seq.

Winton Act processes; (2) the Winton Act is considered a complement to other avenues set forth in the law to deal with employer-employee relations in public school districts and not as the exclusive avenue of communications between certificated employees and school boards; and (3) a "certificated employee council" represents certificated employee organizations; individual certificated employees, who also must be evaluated, are also entitled to represent themselves in their employee relations with their school board. This does not mean that a school board may ignore a request to meet and confer about the guidelines. In my opinion, it is within the ambit of the "meet-and-confer" scope of discussion. But let us keep in mind that the school board must adopt guidelines for the 1972-1973 school year because the Stull Act is operative as of September 1, 1972, and the school board must adopt its guidelines as close to that operative date as possible; it cannot afford to meet and confer interminably on its guidelines. Instead, it should adopt what it has and meet and confer about the regulations. Also, under the Winton Act the school board, while it must make a "conscientious effort" to reach agreement in the "meet-and-confer" process, has the final say on anything before it in that process.

While school boards and certificated employee organizations are under a duty to attempt to reach agreement (but not necessarily to agree), no similar duty is imposed upon the individual employee who represents himself. And this leads us into the fourth question.

May a certificated employee refuse to be evaluated under the Stull Act?

The answer is clearly "no." If the employee refuses to submit to the regularly established evaluation procedures or willfully hinders or obstructs an evaluation, he is insubordinate and subject to dismissal because the law requires such evaluation and the employee is frustrating the school board's effort to comply with the law. The employee's best avenue of protesting the evaluation procedures is probably to use his right to file, along with his own evaluation, a statement expressing his views about the evaluation. Of course, if the evaluation is so negative as to result in dismissal proceedings, he will have ample time to press his viewpoint while defending himself before the "commission on professional competence."

Among other things, though, he probably wishes he could evaluate his administrative superior. This prospect is the subject of question five.

Does the Stull Act contemplate that teachers should evaluate principals or other administrative superiors or that students should evaluate teachers?

The Act really takes no position on these issues. The presumption is that school boards will resolve them in a satisfactory way despite the intense emotionalism they generate. In a word, the Legislature left this up to the local school board by issuing no legal mandates governing such schemes of evaluation as part of the Stull Act.

Another subject not specifically covered in the Act is discussed in connection with question six.

What is the role of the school principal in the evaluation process of the Stull Act?

There is no mention of any particular school administrative official in the Act. In the absence of any such mention, it is fair to assume that the normal functions set forth in state law would control. Under state law, the principal is the head of his school, with supervisory responsibility over all certificated and classified personnel assigned by the superintendent and the school board to the school.[*] Any evaluation system which undercuts the principal's function will surely redound to the detriment of the school district organization.

A practical problem that a principal of a larger school will have is that of personally evaluating all of his teachers and coadministrators. I believe the law will permit an evaluation system which would have school department heads and vice-principals share the evaluation duties of the principal. But, consistent with his responsibilities under state law, the principal should retain final review or veto authority over all evaluations at his school.

May the actual evaluation of certificated employees under the Stull Act be done on a "sampling" basis?

Any reasonable and recognized method of compiling data for evaluation purposes will be considered valid, in my opinion. From a practical standpoint, "sampling" techniques will have to be used, because it would border on the absurd to require the continual presence of an evaluator in a certificated employee's work station. The problem, though, with any "sampling" technique is that it is subject to various defenses, such as: (1) not enough actual evaluations were made; (2) when evaluations were made, the visits to the classroom were too brief; (3) visits to the classroom occurred at an unusual time because the teacher was ill (or the students were agitated due to other campus happenings), so that it was impossible for the teacher to perform up to his usual high standards; and (4) the evaluations on the dates indicated stressed only weak points and entirely neglected the teacher's strong points, so that the evaluations were actually incomplete and therefore invalid.

---

*See Title 5, Administrative Code, Sections 5550-5551 and 5800(m).

What weight will be given by a "professional competence commission" to the "professional judgment" of a principal or one of his teachers who is charged with incompetency under the Stull Act?

As noted earlier, state law contemplates that the principal is the head of his school. There is no reason why his "professional judgment" will not be given considerable weight, providing that such judgment can be substantiated by something more than a vague visceral feeling. In other words, the judgment must be reasonable and based on fact. I believe this will require school administrators and other school people in supervisory positions to articulate clearly their observations and conclusions about persons whom they evaluate. This is not an easy task in the borderline cases. Consequently, it should lead to an improvement in their capacity to observe and analyze their professional brethren.

Are all of the four general criteria of evaluation set forth in the Stull Act of equivalent importance in an actual evaluation?

Generally, yes. The four criteria evidently were designed as an integral unit. They seem to mesh together and provide a foundation on which to base a complete evaluation of any certificated employee. In short, taking all four criteria as a unit, they provide for a total, unified gauge of the overall performance of a certificated employee. Two of the criteria deal with the establishment of standards and procedures, while three concern themselves with assessment of certificated employees under such standards and pursuant to such procedures.

The specific duties of some of the certificated employees being evaluated will tend to provide a natural order of importance among the assessment criteria. For example, a certificated school administrator who has no regular and continuing duties in the classroom will be judged on a different basis from that for a classroom teacher as far as assessing the standards of expected student progress, which is the first of the four criteria listed. The task will be to strike a good balance among the assessment criteria. But, in my opinion, a school district could seek the "ultimate enforcement" through dismissal if the certificated employee being evaluated was deficient in his ratings in either his main work or his collateral duties.

Must an evaluation be totally and completely negative and unsatisfactory to support dismissal for incompetence under the Stull Act?

No, but the extent to which a school district can weed out marginal performers and not merely dismiss absolute losers is going to be the true test of the success of the Stull Act implementation. In other words, if the Act is implemented in a positive manner, its evaluation features should permit more precise determinations of competency than have ever before been possible. This added preciseness should allow for a better showing of proof in the cases of the "marginal incompetents," who are definitely performing below standards, but not abysmally so. If the "commissions on professional competence" and the courts of appeal ultimately require that a certificated employee must be utterly and irredeemably incompetent before the sanction of dismissal will be invoked against him, the Stull Act will have failed in its underlying purpose to upgrade general teaching performance by pegging the practical lowest satisfactory performance too low. This would tend either to cause evaluation programs to be shelved out of a sense of frustration or to subvert the programs by transforming them into projects to prepare virtual indictments against certificated employees who otherwise might have had a reasonable chance for satisfactory growth under a sound evaluation program.

May the evaluations of preceding years be considered in bringing charges of incompetency against a certificated employee under the Stull Act?

Yes, but under Education Code Section 13413, a commission of professional competence is limited in delving back into a certificated employee's record because:

> No testimony shall be given or evidence produced relating to matters which occurred more than four years prior to the date of the filing of the notice (on intention to dismiss by the school board) . . . (and) . . . no decision relating to the dismissal or suspension of any employee shall be made based on charges or evidence of any nature relating to matters occurring more than four years prior to the filing of the notice.

Within those limits, however, past evaluations may be used to show a pattern of conduct culminating in the filing of the notice of intention to dismiss.

Is there a limit to the amount of counseling a school district must provide deficient employees to assist in bringing them up to acceptable standards under the Stull Act?

There are no limits specifically set forth in the law. Therefore, a general rule of "reasonableness" will prevail.

In my opinion, this requirement easily ranks as one of the most crucial aspects of the proper implementation of the Stull Act. The school district has an affirmative duty to provide follow-up counseling and assistance for certificated employees judged to be deficient in their performance. The aim of the counseling and assistance is to remedy such deficiencies. The commission on professional competence will consider most carefully the manner in which help is accorded the deficient performer. Such help will have to be something considerably more than perfunctory advice and will have to be well-documented to satisfy the Commission.

It may well be that in the final analysis the follow-up counseling and advice given pursuant to the evaluation procedures of the Stull Act will prove to be the greatest contribution of the Act to public education in California.

How much follow-up counseling and assistance may be appropriate resolves itself into a question of judgment. Different facts will necessitate different judgments. For example, an evaluation of a teacher could lead to his dismissal; in contrast, an evaluation of a certificated school administrator could lead to his removal as an administrator and reassignment to a classroom teaching position. Not only are the stakes higher for the teacher but the very nature of the employment relationship of a teacher and an administrator vis-à-vis their school district is considerably different. Accordingly, it would be expected that more follow-up counseling and assistance will be necessary in the case of the (ultimate) dismissal of a teacher than in the "demotion" of an administrator.

Does the Stull Act permit any other sanctions besides outright dismissal to be taken against incompetent certificated employees?

No. Generally, California law does not provide for the so-called "intermediate sanction" in dealing with certificated employees who are not performing up to standards. I believe that this is one of the significant weaknesses in the law governing the employment relationship between school districts and certificated employees. In fact, in 1971, while a member of the California State Board of Education's "Ad Hoc Committee on Teacher Tenure," I was among a substantial minority of the committee to file a minority report which specifically recommended that any statute establishing dismissal procedures for certificated employees also should authorize imposition by the dismissal hearing authority of intermediate sanctions against errant or failing certificated employees short of, and as an alternative to, outright dismissal. The reasons substantiating the position of the minority report were:

> Termination of employment is the ultimate sanction exercised by an employer against an errant or failing employee. Under present California law, outright dismissal is the only real sanction which may be imposed upon a teacher whose performance is marginal or less than satisfactory. While there presently is authorization under California law for immediate suspension pending trial of a teacher against whom certain criminal charges have been filed, there are now no provisions in state law which contemplate, for cause, the teacher's suspension without pay, or his forfeiture of tenure rights, or his reversion to intern status, or his loss of a future salary increase, or other intermediate action taken as a reasonable alternative to outright dismissal under all the facts of an individual case. In sum, the only real action which may be imposed against a teacher under today's state law is outright dismissal.
>
> We believe that this rigidity inhibits resort by school boards to dismissal proceedings and places effective constraints on the Superior Court in rendering verdicts in marginal teacher dismissal cases under present law. We believe that if options to impose intermediate disciplinary or personnel management sanctions short of outright dismissal also were available to the teacher dismissal hearing authority (as they are in most civil service and private sector employee dismissal proceedings), many problems relating to teacher tenure which one hears people complain about today would be solved.

"Intermediate sanctions" may be a subject of amendment to the Stull Act in years to come.

What kind of record keeping will be required by a person who evaluates a certificated employee in order to sustain the employee's dismissal or other discipline on the basis of incompetency under the Stull Act?

Record keeping in the public personnel field is an art in itself. On the one hand, the evaluator must keep sufficient records with which to substantiate his recommendations, and on the other hand, he must guard against his records being used against him in an attempt to show that he was either deficient in the discharge of his responsibilities or biased against the person who was the subject of his evaluations. There seems to be a gray area between one end of the field, where the charge is that "I wasn't evaluated enough," and the other end of the field, where the charge is that "I was evaluated so much it amounted to open harassment." The frequency and adequacy of evaluations go hand in hand with record keeping, because together they give substance to support any dismissal action based upon lack of competency.

There are several practical techniques to observe in keeping records in the evaluation process. The major ones are: (a) record observable and verifiable facts and not opinions, where at all possible (don't say, "William was unkempt and obviously intoxicated." Instead record the specific kinds of behavior you noted which lead you to those conclusions about William.); (b) record strengths as well as deficiencies; (c) be thorough and not haphazard in your recordings; (d) use the same system of recording for all personnel who are in similar situations; (e) use straightforward language with as few adjectives as possible; (f) be candid and forthright--don't use the recordings as a "secret" document, but review them with the person being evaluated and let him read it. In this connection, it is important that the recordings be viewed by others as a memory aid and a positive force in helping certificated employees to improve their performance. The regard which certificated employees have for the recordings is very important. If the effort of recording takes on the aura of preparing "secret indictments," the impact upon morale in the school staff is devastating. Finally, (g) don't display the recordings openly in public or jot things down in plain view of your staff or you'll soon develop a reputation for being a bit of an ass.

Is an evaluation which the certificated employee believes to be unfair or incorrect a proper subject of an employee grievance filed under the provisions of a locally adopted school district grievance procedure?

The Winton Act does not require school districts in California to adopt employee grievance procedures. However, many school districts (especially urban districts with large staffs) have adopted grievance procedures as the best form under all the circumstances to deal with employee complaints in a fair, orderly, and efficient manner. If a school district has a grievance procedure, then the answer would hinge on whether the evaluation will form the basis of a dismissal action. If such is the case, the proper forum to contest the evaluation is through a hearing before the commission on professional competence. If, however, the evaluation will not be used at the time it is made to buttress a dismissal case, the grievance procedure would appear to be the right vehicle to test the validity of the evaluation. In any event, the employee has two shots at the evaluation, in that he may file a written objection to it as specifically permitted by the Stull Act and he may either file a grievance procedure complaint or defend against the evaluation before the commission on professional competence.

As a practical matter, when and under what circumstances would it be expected that the courts will begin interpreting the Stull Act?

I suppose this is another way of asking when the dismissal lawsuits or administrative hearings will begin and what will be the issues. Some local teacher organizations throughout our state have taken the position that the Stull Act's requirement of teacher involvement in the development of evaluation procedures must be read together with the Winton Act and that the actual development of such evaluation procedures is exclusively subject to the organizational "meet-and-confer" processes. This contention will be raised in the courts, either as a separate lawsuit instituted by a teachers' organization because of its "organizational security" ramifications, or as one allegation among others in a lawsuit testing the validity of a dismissal brought on behalf of a teacher actually proposed for dismissal pursuant to an unsatisfactory evaluation under the Stull Act. But this issue will test only the procedure followed in developing the evaluation program. The first real test of the Stull Act evaluation programs most likely will occur in the spring of 1973, during the period of probationary teacher nonreemployment cases under Education Code Section 13443.

The state hearing officers of the California State Office of Administrative Hearings are required by law to conduct probationary teacher nonreemployment hearings. The state hearing officers who will conduct probationary teacher nonreemployment cases next spring will, under the Stull Act, consider several issues in cases involving the competency of a neophyte teacher whom the school district accuses of not being able to perform satisfactorily. And the burden of proving these issues will fall upon the school district. The issues may be sketchily described by the following questions:

1. Did the school district adopt evaluation guidelines required by the Stull Act in sufficient time to permit adequate evaluation of the probationary teacher for the 1972-1973 school year?

2. Are these guidelines reasonable under the circumstances prevailing in the school district in which the guidelines are applied?

3. Are these guidelines written in reasonably clear and precise language so that the persons doing the evaluating and the persons being evaluated can be expected to understand the guidelines' intent, purpose, and effect?

4. Were the guidelines made known to the probationary teacher in sufficient time to permit the teacher to act upon them?

5. Were the guidelines actually observed by school district evaluators?

6. Was adequate follow-up counseling for deficient performers actually provided?

7. Was the rate of progress, if any, measured by school district evaluators in the follow-up counseling process?

8. What is the prognosis of adequate follow-up counseling?

In the hearing of a probationary-teacher-nonreemployment case involving the capacity of the teacher to perform, all of these general questions and a myriad of other questions flowing therefrom will have to be answered in a satisfactory manner by the school district. The thrust of the inquiry may be summed up in this way:

A school district may not hold a teacher to standards which are not properly adopted, not reasonable in substance and scope, not clearly stated, and not sufficiently promulgated, and where no real effort has been made to cure the defects of performance noted, as required by law.

In a sense, there is nothing more than a rudimentary form of "due process" imposed upon the interpretation of a teacher performance evaluation statute. But it will be a rude awakening to any school district that has adopted a cavalier or perfunctory attitude toward implementation of the evaluation section of the Stull Act under the comfortable, but entirely false, impression that the Act does not affect probationary teachers.

Finally, in future cases of dismissal of permanent employees and probationary employees for cause pursuant to Education Code Sections 13403 and 13442, the first confrontation in the "ultimate enforcement" of the Stull Act evaluation program will occur when the written notice of incompetency is delivered to an employee under Education Code Section 13407, which must be at least 90 days prior to filing the notice of intent to dismiss for reasons of incompetent performance. This written notice of incompetency must specify the nature of the incompetency

> with such specific instances of behavior and with such particularity as to furnish the employee an opportunity to correct his faults and overcome the grounds for such charge (of incompetent performance). . . .

As a practical matter, for dismissal of permanent employees and probationary employees for cause, such written notices of incompetency will not be sent until fall, 1973, at the earliest because of Education Code Section 13405, which generally prohibits notices of intention to dismiss from being given by school districts in professional incompetence cases between May 15 and September 15 in any year.

These are the legal implications of evaluation under the Stull Act. The impact of the Act upon public education in California will depend in large measure on the extent to which these legal implications are understood and dealt with by school boards, school administrators, and teachers in the next few years.

Appendix A

A "BIRD'S-EYE VIEW" OF THE LAW GOVERNING PROFESSIONAL EVALUATION AND DISMISSAL OF PUBLIC SCHOOL CERTIFICATED EMPLOYEES IN CALIFORNIA

State law governs the professional evaluation and dismissal of public school certificated employees in California. Any regulations relating to these two subjects issued by the state or local school district boards of education must comport with state law. The 1971 Regular Session of the California Legislature significantly changed such state law by Statutes 1971, Chapter 361, popularly referred to as the "Stull Act." Local school district boards of education may adopt regulations to implement the Stull Act.

The Act is divided into two major divisions, one covering the professional evaluation of certificated employees and the other specifying the legal grounds and procedures for dismissing such employees. The Act applies to _all_ certificated employees, from district superintendents to credentialed teacher assistants, in the public elementary and secondary school districts in California. It does _not_ apply to certificated employees of community college districts. In brief outline form, the Stull Act may be summarized as follows:

I. _Part Governing Professional Evaluation of Certificated Employees_

1. _Effective Date_. This part, which adds new Secs. 13485-13489 to Ed. Code, becomes effective on September 1, 1972 by virtue of enactment of Statutes 1972, Chapter 10.

2. _School Board Must Adopt Written Guidelines_. Each school board must adopt a uniform set of written objective evaluation guidelines for use in evaluating professional competency of its certificated personnel.

3. Contents of School Board's Written Guidelines. The written evaluation guidelines for certificated employees adopted by each school board must include, at least:

a. Establishment of standards of expected student progress in each area of study and of techniques for assessment of that progress.

b. Assessment of certificated personnel competence as it relates to established standards.

c. Assessment of other duties normally required to be performed by certificated employees as adjunct to their regular assignments.

d. Establishment of procedures and techniques for ascertaining that certificated employee is maintaining proper control and is preserving suitable learning environment.

4. Follow-Up Counseling for Deficient Certificated Employees Required. Each school board must provide for a follow-up counseling of certificated employees whose professional competency as judged under the school board's written guidelines is not up to standard and must also provide assistance for such employee to upgrade his performance.

5. Evaluations Must Be in Writing and Guidelines Must Be Published. Evaluations must be in writing and transmitted to certificated employee not later than 60 days prior to end of school year in which evaluation takes place, commencing with the 1971-1972 school year; a copy of the school board's written guidelines also should be provided each certificated employee in advance of his written evaluation.

6. Certificated Employee Evaluated Has Right to Rebuttal. Certificated employees evaluated must be given opportunity to append written statement of their views about evaluation to evaluation report and such statement shall be permanent part of their personnel file.

7. School Board Must Seek Advice. In developing and adopting written evaluation guidelines, school board must avail itself of advice of certificated instructional personnel.

II. Part Governing Dismissal from School District Employment of Certificated Employees

1. Effective Date. This part, which extensively amends Ed. Code Secs. 13403 et seq., relating to dismissal of certificated employees, does not become effective until 61 days after final adjournment of 1972 Regular Session of California Legislature. Until then, dismissal of certificated employees may be accomplished only under present state law.

2. Dismissal Hearings. All dismissal hearings of permanent and probationary certificated employees will be conducted out of court as administrative hearings in accord with State Administrative Procedure Act and must commence within 60 days of an employee's demand for a hearing.

3. Two Kinds of Dismissal Hearings and Eleven Legal Grounds for Dismissal. All dismissal hearings of certificated employees will be conducted by either a state hearing officer or a local "Commission on Professional Competence." The eleven grounds for dismissal are not changed by the Stull Act. The two kinds of hearings and eleven grounds are:

a. State hearing officer shall conduct hearing in cases of allegations concerning criminal syndicalism, physical or mental condition unfitting one to teach, conviction of felony or any crime involving moral turpitude, knowing membership in the Communist Party, indoctrination of pupils in communism on school grounds, and refusal to appear before, or answer questions concerning affiliation with Communist Party or any organization advocating forceful or violent overthrow of U. S. Government propounded by, school board or committee of the Congress or California Legislature.

b. "Commission" shall conduct hearing in cases of allegations concerning immoral or unprofessional conduct, dishonesty, incompetency, evident unfitness for service, and persistent violation of state or local laws or regulations prescribed for government of public schools.

4. Effect of Decisions. Decision of State Hearing Office and majority decision of "Commission" are binding on school board and employee but either may appeal to Superior Court, where such appeal is entitled to precedence on trial calendar.

5. Costs of Hearings. If employee is ordered to be dismissed, employee and school board split hearing costs and each pays own attorney fees; if employee is ordered reinstated, school board pays all costs, plus "reasonable attorney fees incurred by the employee."

6. Membership of "Commission." Composed of three members, one selected by school board and one by employee (both must have five years experience in educational function of employee). One must be state hearing officer, who also acts as chairman and protects "legal rights" of employee. If parties have not selected commissioners at least seven days prior to date of hearing, County Board of Education shall make selection for them.

7. Nonreemployment of Probationary Teachers Not Affected. Ed. Code 13443 will continue to govern hearings of probationary teachers notified prior to March 15 that they will not be reemployed for ensuing year.

8. Demotion or Reassignment of School Administrators Not Affected. General employment contract law will continue to govern demotion or reassignment of school administrators notified under Ed. Code 13443.6 prior to March 15

that, while they will be retained as direct employees, they will not continue to hold same administrative position.

Thomas A. Shannon, October 31, 1971; Revised, May 25, 1972.

DISCUSSION OF PAPER BY THOMAS A. SHANNON

JACK H. FRIEDENTHAL
Stanford University

Speaking as one who is not involved directly in schools of education or in the teaching of children below the college level, I was appalled when I read the Stull Act because of the implications I see in the Act and in the lack of interrelationship between its two parts--that is, the new evaluation procedures and the problems of dismissal of teachers. As Dr. Shannon pointed out, everything done in terms of the evaluations is looking toward the problem of dismissal, of getting ready at the time of the evaluation to justify the dismissal of an incompetent teacher whether tenured or not. That is a very upsetting thing. I would hope, instead, that the Stull Act would be used primarily to help improve teacher performance, by telling teachers what their deficiencies are and working with them in a realistic way to improve performance.

The reason I think the emphasis is misplaced is that, as most of you know, very few teachers are dismissed for incompetence and very few dismissal hearings are held. This situation isn't going to change very much through the use of the Stull Act procedures. If the procedures could be utilized to improve the performance of those who are going to be retained, there would be a much more positive benefit. Unfortunately, lawyers are going to get hold of these procedures and make them basically into instruments of legal contention. So, as Dr. Shannon points out, you should all be aware of those legal implications.

Speaking strictly about what the relationship is between the two parts of the Act--the part that deals with the new evaluation techniques and requirements and the part concerning the filing of charges and the holding of hearings, I would like to make a few comments. As most of you undoubtedly know, anybody can file charges before a school board charging a teacher with incompetence, as long as the charge is signed under oath. That is provided for under Section 13404 of the Education Code. The board doesn't have to act on such claims, but a person can file a complaint if he feels that a particular teacher is not doing the job he would like to see that teacher do.

This raises the question of who is going to see the evaluations. Dr. Shannon points out that principals who are going to be doing the evaluations ought not to do them out in the open, where people not affected can see them, and I agree with that. But who will be on the evaluating committees? Are there going to be parents involved, or students? I hope not the latter--we've had enough of student evaluations on an ad hoc basis! I happen to be dead set against that for many reasons. Nevertheless, the uncertainty about who the evaluators will be merely increases the problem. The Act says nothing about the confidentiality of the reports. I am seriously concerned about this--there are other areas in the Education Code that call for confidentiality; Section 13123, for example, has a requirement of confidentiality with respect to certificate hearings, the State Board of Education has to keep things confidential among members of the staff, and so on. They are not allowed to reveal what goes on at hearings, and indeed that is considered so important that it is a misdemeanor for someone to reveal such proceedings improperly. So I am concerned with the fact that confidentiality is not written into the Stull Act. Let us assume that in some evaluation proceedings some deficiencies of a particular teacher are noted by an honest principal, not for the purpose of dismissal in the long run, but for the purpose of assisting the teacher to improve his performance. The next thing you know, an unhappy set of parents gets hold of this evaluation and charges are filed--in the newspapers, not just with the board--and the school board is suddenly under unwarranted pressure to fire the teacher.

I wish we could read into the Act a provision that the evaluations of individual teachers would be thrown away after consultation with the teacher, because I feel their primary purpose should be to improve teacher performance, not to make a record so that some time in the future we can go back and say: "Hey, he had a dirty bulletin board the day I was in his classroom." I'm very sensitive about this because in the last case I had, the firing of a probationary teacher was upheld on the basis of a dirty bulletin board. The case upset me quite a bit--there were 60 specifications, and 58 dealt with room environment and trivial violations that I think no teacher could have remedied satisfactorily if someone was "out to get him." Under the Stull Act, the evaluations are part of the permanent record, unfortunately, and as Dr. Shannon pointed out, there are four years in which one can go back and look at these evaluations, taking casual matters that were put down by an honest evaluator and building them up into a "case" for dismissal. It can happen.

Another thing that bothers me is that the Stull Act contains no provisions for recording the results of consultations. Under the Act if there is an unsatisfactory performance the evaluator is specifically required to so inform the teacher, and then there are to be conferences and the like about these matters. Yet nothing is said about putting the results of these conferences--what the teacher said, etc.--into the record. As long as detailed records are to be kept, these matters should be recorded too. There ought to be follow-ups from the consultations, not simply an evaluation every year about how a teacher is doing or what his classroom looks like, but what the teacher has in fact done to remedy deficiencies; I think a burden will be placed upon school administrators and evaluators to make and record such follow-up evaluations, if for no other reason that that they are going to be raised as defenses by teachers if they in fact get to court.

Under the Stull Act a teacher can file a response to an evaluation. If the teacher feels the evaluation is unfair he can say so in writing, and that, together with the evaluation, becomes a permanent part of his file. One of the difficulties that I find with this provision is that it is unclear how it will be used in a courtroom or a proceeding before a professional competence committee. Suppose, for example, that no response is filed. Does that mean that the teacher admits everything that is said against him? Most teachers aren't going to be looking toward dismissal. I hope that every teacher will get some comments which would help him improve his performance. Yet if he doesn't agree with them, is he really bound to say so or, so to speak, forever be bound? This must be clarified, or every teacher will need to have a lawyer to respond to charges for fear that otherwise the charges will be considered true.

Another problem, from a legal standpoint, is whether evaluations under the Stull Act can alone meet the burden of proof necessary for a dismissal. The Education Code provides for the use of regularly kept records in dismissal hearings. The new section with respect to the hearing officer says that not only can testimony on evidence be taken, but that the records regularly kept within the past four years may be used. But will it be sufficient simply to submit these evaluations? Maybe the evaluator is no longer available to testify; will the records alone be enough to uphold a charge of incompetency? We don't know many answers as to what the burden of proof is, how far it extends, and on whom it rests; these are questions that are still not completely decided and certainly are not dealt with in the Education Code.

The disproving of charges by the teacher raises some interesting questions. Under the Stull Act, if the performance is unsatisfactory the teacher must be told so in his evaluation. Of course, he must be told of his deficiencies at some point before charges can be filed, but suppose that these do not appear in the formal evaluation. There is some question as to whether an administrator can give an evaluation praising the teacher on the one hand and then suddenly tell him, independent of that evaluation and at some later point in time, that he is not doing a good job. It thus appears that the two entirely separate provisions of the Act do not mesh at all. It is unfortunate that they don't, because it leaves the way open for the kind of situation I have just outlined, where the teacher is misled as to the quality of his work.

Now let us suppose the formal evaluation is not made at all, and that a teacher is simply told in the middle of a semester that he is not doing a good job and within 90 days charges will be filed to dismiss him. I don't see the lack of an evaluation as a defense. Dr. Shannon indicates that it will be, and as a practical matter that is the way it may be held, but no one can be certain at this time.

Another problem that I see stems from the notion that in evaluating different people, the evaluator must be consistent. I agree that evaluations should be made according to a consistent procedure, but this raises a question. Suppose someone is negatively evaluated, that charges are in fact filed, and that he has a hearing before the professional competency committee. May that teacher subpoena the evaluations of all the other teachers in the school? He could say, for example: "Look, this is what my evaluator said about another teacher--he's got a dirty bulletin board too, so why don't they fire him?" He could just tick off the list of deficiencies and ask: "Are the decisions as to various teachers really consistent?"

Finally, are the standards that will be used for the evaluation system the same standards by which a teacher will be fired? Dr. Shannon indicated there might be a difference and that the courts might say that only the absolutely lowest group can be fired. In other words, a higher standard might be set for a school district than the courts are willing to enforce. Remember, the Act has two entirely different parts--one talking solely about discharges for incompetency and the other saying that districts may set standards for pupil progress and the like. This doesn't necessarily mean that the two will mesh, and that administrators can rely particularly on the standards they set. If standards are set and people meet them, they probably cannot be discharged; but standards may be set which people don't meet but which still will not be enforced by the courts.

In summary, I look forward to some changes in the Act. What I would hope for primarily is enactment of a provision restricting the use of evaluations and protecting their confidentiality, plus some language that would mesh the two parts of the Act to ensure that the standards set by the school district will be the standards that will be enforced by the professional associations and the courts.

PETER T. GALIANO
California Teachers Association

As a former school board member I am always bemused by the comments made by administrators in speaking of evaluating teachers' performance. I am also bemused as an attorney who has represented teachers in dismissal proceedings. And that bemusement continues, particularly when Dr. Friedenthal mentions cases such as dirty bulletin boards. I think Dr. Friedenthal's dirty bulletin board example compares with the one in my own experience with the teacher who was alleged to have failed to arrange the furniture properly in his room. The only piece of furniture in his room, which was there when the building was built, was a table bolted to the floor! The major charge against another teacher I represented was that the shades along the windows were not even on Friday afternoons. That was the specification--on Friday afternoons. I never could understand what that meant.

As a school board member, I have had a limited experience with observing administrators as they evaluate personnel. I've never quite got away from the conviction that a school principal's major job is to kick the tanbark off the sidewalk every day, and I've never been quite convinced that the school administrator understands and realizes that his position is directed to the development of people and not to the management of things. There are expressions in the Stull Act which indicate this objective.

I think that, in addition to the amendments and clarifications that Dr. Friedenthal has indicated are necessary, perhaps a most immediate measure should be to require every administrator in California to read Mr. Shannon's paper. That would be a good beginning, because his paper does emphasize this primary responsibility of administrators in the development of people. I have to agree that there is an unfortunate and inescapable correlation between the dismissal provisions and the evaluation procedures of the Stull Act. When you consider this, as Mr. Shannon has mentioned, the competency panel can make only one of two decisions, that is, dismissal or nondismissal. As you can see, that conclusion is reinforced where the evaluation procedures are so closely connected with the dismissal, and I think this is a deficiency in the legislation.

There are other deficiencies, of course, that are not directly related to the evaluation process. Some of these will offer some constitutional problems. I refer briefly for the moment, for example, to the charge of criminal syndicalism, an anachronism which is in the Stull Act. I don't recall any instance where this has come up in an actual hearing, but I think it does demonstrate some antiquity in the thinking involved in this legislation.

I want to quote a statement made by Assemblyman Stull:

> I believe that Assembly Bill 293 provides an answer to the school districts' and the citizens' demands that incompetent teachers be weeded out of the educational ranks, and that the young people attending our elementary and secondary schools be provided with the best qualified teaching personnel available.

Again we have that correlation between evaluation and the weeding out of personnel.

I heard the Executive Secretary for the California School Board Association state to the Assembly Committee on Education that the school boards want to take advantage of the buyer's market, and that this would be one of the purposes of legislation amending the tenure act. Now, as I am sure you are all aware, the Stull Act is obviously a compromise. The initial attempt in the Legislature last year and for some time has been to eliminate the very concept of tenure with respect to public school employment. That objective failed of realization and we wound up with the Stull Act. This is the act under which we will be operating for perhaps the next two years.

I might mention that the California Teachers Association maintains an extensive legal service program available to its members, particularly to tenured teachers. And as Mr. Shannon and Dr. Friedenthal have indicated, the provisions of the Stull Act are under analysis by defense attorneys for teachers.

In the case of the California Teachers Association, there are 25 law firms located throughout the state that will be ready, willing, and able to represent teachers in dismissal proceedings. You may be sure they will exhaust every opportunity given to them under the new discovery procedures. That does not mean that there is a particular "mind-set" on the part of our attorneys in the sense that they intend to oppose you administrators in every respect. As attorneys, we are not really interested in whether the teacher is competent or not; that is a judgment for professional educators to make. The attorney's position is to see that the procedures are in fact followed and that every due process defense is made available to the teacher in his attempt to retain his position as a teacher in the school district.

Regarding the uses to which an evaluation will be put, let me remark that there are school districts in California that ask the teachers to come in individually and discuss evaluation guidelines and objectives, both in terms of the broad concepts established by the school board and the concepts established by the school principal and the teacher. And after they have arrived at some apparent consensus the principal asks the teacher to sign the document with such words as "I acquiesce," or "I concur," or "I agree." The principal is surprised when the teacher hesitates, and the school board and the school superintendent are surprised when we advise them that we have told our teachers not to sign such a document. I am not quite sure what the utility of that signature is.

Mr. Shannon indicated that the adoption of standards by the school board is not a proper subject of the meet-and-confer process under the Winton Act. In other words, there is no requirement on the part of the school board and the Certificated Employees Council to reach some kind of agreement, or at least a proposal in form which could be submitted to the school board for adoption. I do not agree with that. I feel that probably the most important subjects with respect to the professional life of an educator are his performance and the performance of his students. If these don't constitute terms and conditions of work, I'm afraid I would not be able to recognize a term and condition of work in any context. Yet we are told that this is not subject to the meet-and-confer process. But let us put that aside for the moment--it will be tested in the courts. We have had some satisfactory results in a somewhat different area with respect to the obligation of school boards to meet and confer.

It is the school board's ultimate authority and responsibility to adopt standards of performance for certificated personnel. If that is true, then there is no legal utility (or any other kind of utility that I can understand) in asking a teacher to sign a document in which he acquiesces in certain standards of performance.

We are speaking in terms of performance in the context of evaluating the performance of a professional. If I had some criticism to make of Mr. Shannon's performance as an attorney, that question would be decided by attorneys, and only by attorneys. And we wouldn't let you know about that.

We would keep that in the profession. I think it is outrageous to ask a school teacher (perhaps a terrified teacher, a newly employed probationer or perhaps a person who has devoted many years of his life to his classroom without taking any notice of these legalities with which we are now concerned) to come into a principal's office, alone, and to discuss at those two levels of sophistication the evaluation of his performance on a somewhat formalized basis, without the comfort, support, and strength of his organization. I do not understand this. From my own understanding of establishing standards in terms of professional performance, I wonder why a teacher's standard of performance must be measurable and understandable and explicable only in terms of standards established for all teachers in that area. But if that is true (and I think Mr. Shannon has indicated that there must be some kind of uniformity with respect to the performance of teachers in a given area), then why is it that those teachers cannot in concert represent themselves in concert, develop standards for performance among the teachers, and, in concert, meet with the school principal? I think this would be much more reasonable and rational in terms of establishing the validity of standards for the performance of professional personnel.

RAOUL TEILHET
California Federation of Teachers, AFL-CIO

I have heard Dr. Shannon speak on many occasions and I'm very familiar with the law and its background. The California Federation of Teachers bitterly opposed the adoption of AB 293; we were almost alone in the field of education in this opposition. But it should be clearly understood that the California Federation of Teachers fully supports evaluation, accountability, general goals for public education, specific objectives for each teacher in the classroom, some mechanism to see that they are accomplished, and ways and means to improve instruction to benefit the institution, the students, and the community. The difficulty here for us is that we are a teachers' advocate organization. We are a union. We represent our members, we defend teachers, and we advocate collective bargaining and arbitration proceedings to do exactly the same thing that is jury-rigged into the Stull Act. But at the same time that we are trying to improve the quality of the service we render, we feel a responsibility to the institution itself.

I imagine that millions of dollars in man-(and woman-) hours have been lost already in the State of California in talking about the Stull Act. I consider it to be one of the most irrelevant subjects to be confronted by public education and professionals within the institution in a long, long time. I admire its political success in getting us all to chase our tails around like dogs, spending a great deal of time and energy believing that we are going to accomplish something of substance, while those who are really making progress in a different direction are proceeding without any opposition.

As a thesis I want to suggest that there is absolutely nothing in the Stull Act that has anything to do with improving the quality of instruction in California public education that did not exist prior to its enactment. The old law was bad enough; as a teacher I was embarrassed by it. I felt it was a degrading kind of statute. Its only saving grace was that there were only three or four people in the entire State of California in each given year who read it. It was a bad law. It was written in response to a more terrifying experience, and that was the unilateral and capricious authority of California school administrators and school boards to fire teachers prior to its implementation in the late 1930s, without something called due process or substantive due process. The specified reasons for dismissing teachers were for the most part unconstitutional. Now I thought that the Legislature might have had a golden opportunity and that they were talking about substantial reform of this law to eliminate unprofessional conduct, which hasn't been used in California with any success since Owens v. Lassen County, because you can't define what "professional" is, so it is hard to define what "unprofessional conduct" is. The "evident unfitness to serve" cause has been used as a catch-all for just about everything that was originally designed to deal with mental incompetency and emotional instability in the classroom, criminal syndicalism, membership in the Communist Party, and a whole host of other violations of California penal codes, health codes, and other kinds of codes alleged as grounds for revocation of credential and dismissal. If the accountability concept of the Stull Act were used in implementing the old tenure laws that specify reasons for dismissal which have not been altered one whit, everyone that I have ever known casually and otherwise in public education would be subject to immediate dismissal and suspension by their school boards. Most of these laws are unconstitutional and the rest of them invade the privacy of the home, one's morals, and how one lives one's life. But I didn't see Assemblyman Stull or any other brave legislator step forward and demand that these laws be removed.

I shan't share with you my file of atrocity stories for dismissing teachers, but I will say that people who laugh at such things as teachers being fired for dirty bulletin boards have never been fired, much less for dirty bulletin boards. We consider it to be a very serious type of experience. I feel that the hidden purpose of the Stull Act is to keep our minds off the fact that we are an underfunded institution, and I believe that this is the outcome of a political attack that started in the early 1960s because of the Free Speech Movement, or campus disorder, in California. Assemblyman Stull was chairman of the Assembly Subcommittee on Campus Disorders at that time. I submit that he has never psychologically recovered. The reasoning behind the Stull Act seems to be that if we get rid of inefficiency in the classroom and in administration we will find we can probably educate children much better with less money than we have right now.

Section 13407 of the old tenure law (and in the new one nothing has substantially changed) has always provided administrators with the right to identify incompetency in a classroom and to tell the teacher that, unless remedial change occurs within 90 days, a resolution would be recommended to the school board for termination of employment. This provision has not changed since it was originally adopted, and it hasn't changed now. School boards have always had the authority to identify student achievement as a mechanism for evaluating teachers. They have always used classroom learning environment, and they have always used out-of-class considerations to evaluate teachers (in other words, watching school buses, chaperoning dances, holding hands in the faculty lounge, or things of that nature).

I happen to support all of these evaluation measures. I think that student achievement, classroom learning environment, and general conduct as an employee at the work site are legitimate and valid reasons to be evaluated for any employee. Too many teachers have never been willing to admit to the fact that they are employees, and I think teachers have to recognize this and to realize that there _is_ an employer who has a right to make certain demands. As employees, of course, teachers have the right to collectively do certain things on their own behalf, and to negotiate protections and environments that they need to accomplish the goals as defined by the employer. As an employee you always have the right, whether in public employment or in private employment, to defend yourself, through a union in private employment or through an association or union in a public institution. You have the right in both instances to demand due process of law, through arbitration, an administrative hearing, or a hearing in a court of law, from somebody who is impartial. I think that is part of the American tradition, the kind of thing we are presumed to teach in the classroom. It would be novel if some of these attitudes carried over into our experience as employees.

The problem with educational management is that it has never been willing to admit that it is management. It wants to be called a principal, an educational leader, a catalyst, or some other nonsense out of the space age. I recall that Tom Wood, when he worked for the Elementary Principals' Association, said that educational administrators are "mid-management." It is a proud definition and a proud role; let's do it efficiently and competently and represent the institution that pays us. Now I know a lot of principals who don't like to hear this kind of talk. Their problem is that they must wear two hats. They are managing a very expensive public institution, on the one hand, and on the other hand they are educational leaders and they are going to evaluate teachers. The error is that most of them are not competent to do so. I would be terrified if I were ever asked to be a principal of a comprehensive secondary high school and had to implement the Stull Act.

Now let me deal with some legal responses to the Act. I think that Dr. Galiano was most accurate. We have the very finest lawyers in the State of California and they have had a wide experience in defending teachers. We will demand that evaluations be, as Dr. Shannon indicated, objective, precise, and substantial in nature, and that the goals be understood. We will subpoena all the documents that we can get to defend our teachers. We do not view a seventh-grade social studies teacher in a vacuum. Where there are other seventh-grade social studies teachers and other teachers in the school district we expect some kind of measurement of fairness and relativity to be involved before value judgments are made. This business of confidentiality is an old saw used by management to hack up employees since the Year One. We work for a public, political institution. Anyone who doesn't understand that just doesn't deal with public education. School board members are elected. They run on a partisan basis. They are conservatives or they are liberals and they run on that basis. They select their administrators that way, and they adopt policy on that basis. And since the schools are a public institution I think the people have the right to know about all their business. If it benefits any of our clients to "go public" with an issue and to demand a public hearing, then that is exactly what we will do.

We strongly believe that there is a correlation or relationship between the Winton Act and the Stull Act. We think most school districts in California violate the Winton Act by not meeting and conferring, or as we call it, collective bargaining. They must abide by their own policy by meeting and conferring on the procedures by which the Stull Act is going to be implemented. I think it is more than appropriate that Dr. Shannon has stressed that position, because within approximately ten days we will file suit in the San Diego Unified School District as one of three test cases we are using in California on that question. He may have to modify his paper the next time we go around, but we consider Tom Shannon to be the very best school lawyer in the State of California on the side of management, so there is a good chance that we might have to change our position the next time around!

The California Federation of Teachers supported SB 696 for community college teachers, a much more enlightened law which employs certain generalized language on the use of sanctions other than dismissal. There might, for instance, be other options open to a school board, such as a year's suspension or a year's leave of absence, or maybe a year's remedial training program to shore up some weaknesses in classroom instruction--in short, anything other than dismissal, particularly in the field of morals. When we find teachers doing an excellent job but who are suffering from a social disease such as alcoholism (and it should be viewed that way), the board should have the option to permit them a year's time to recover, to take care of the disease just as if it were a communicable disease. But right now the board has to retain such a person totally or to fire him. I think we have moved to a period of enlightened relationships where we can start to make these kinds of recommendations to our own constituencies and membership.

I would like to make some recommendations to you. First, when you do go into evaluation spend some money on it, for once in your lives. Provide released time and training time and remedial in-service time as a natural extension of good evaluation. I just want to tell you that every time we get around to the cost of the evaluation, I agree with Dr. Shannon that it is needed, but don't say that we are going to support it. However, our rhetoric dissipates in the political realities of school board budgets. We believe that one good thing the Stull Act has done is to help us in the probationary dismissal cases. We think we have a lever in defining what "cause" means. It used to be that the board had total discretion in determining cause and sufficiency of cause in dismissing the probationary teacher. We are going to demand the same criteria, the same standards, and the same basis for evaluation and dismissal for probationary teachers as we now have for permanent teachers. One of our basic lines of attack from the very beginning has been to ask you (the evaluator), if you were in the classroom one day and observed, did you come back the next day, and the third day, and observe the same class and the same circumstances? And did you come in the day before you observed to discuss with the teacher what he or she was going to do the next day, to look over the lesson plan and to have some understanding of what was going on in the classroom? If the answer is "yes," it is probable that a valid judgment was made about the teacher. But most evaluators come into a classroom once or twice a year and that's it. To sum up, it costs time and money to do a proper evaluation.

Second, I recommend that you spend some money on in-service training. Take some teachers out of the classroom and let them observe other teachers. Set up training programs, have lectures, have demonstration people invited in, and give the teachers relief time. Every other industry does this to improve the quality of the services they render. Spend some money on improving the quality of the program that you presume to evaluate.

Third, avoid overkill. For you who are involved in this business, become more secure with your own judgments. I am tired of hearing 85 pages, 95 reasons, and 432 specific reasons to unload a teacher who can't teach reading or has a personality problem and can't get along with children. Go in there with your facts, discipline yourself to them, don't write down everything that comes through the garbage disposal unit, and you will find that hearings will be shorter, more rational, render a better judgment, cost less money, and put you in a better position with your boss. Those of you who are the bosses of people who are making these kinds of decisions should view with some suspicion all cases that are presented and look at them very carefully for hidden reasons. In more cases than one you will find that you are dealing with a personality conflict, an administrator bias, or some other socio-political reason and that your site administrator, being a human being, has subconsciously used his power to bring on a conscious attack for, say, dirty bulletin boards. And when you see those kinds of charges go back and investigate them from a managerial standpoint, with a sense of responsibility.

I suggest also that you adequately prepare your county counsel. Most of them do not understand public education--the relationships among teachers, counselors, students, school boards, deputy superintendents, consultants, school nurses, psychometrists, psychologists, parents, PTA's, booster clubs, and so on. They need to be briefed, they need to develop some expertise. They get in a lot of trouble in these hearings and then you go back to the board and say, "We had a lousy attorney." We don't think that is absolutely the circumstance.

I now come to some concluding points. One of the possible blessings of the Stull Act is that, as a by-product of all this waste of time, it has elevated the awareness of certificated public employees in California as to just what a bad law they had had and how much worse it has become due to their acquiescence in not paying attention to what is going on. I believe this will lead to very positive reform of a much different nature from what either of the previous speakers might have had in mind. I believe we'll move to collective action and written contracts and grievance procedures within the institution, so that we'll eventually avoid taking our problems before the public and have administrative hearings and Superior Court proceedings. Our organization will demand due process. We expect that those of you involved in management will also become more responsive to that procedure.

Is the Act going to improve instruction? I leave you with a serious reminder. The only way I know that we can improve education using the things we know about teaching in a classroom is that there are certain behaviors on the part of a teacher that block learning and are bad for children--that is, when they are embarrassed, humiliated, or attacked, when information is distorted or improperly brought before them, or when they are treated in a manner less than humane. We also know that there are good things that happen in the classroom when the reverse of this situation occurs--when there is a positive environment where a child feels comfortable and is interested and can proceed. There is no such thing as a standard for every teacher; we are extremely different kinds of people, of different ages and different social and political backgrounds. Some of us are superteachers and some of us are weak, some of us are older or younger, some are tired and some are strong, some have an overload and others a small load. All these variables affect our performance. We must recognize and accept all these variables and come up with enough money to do some research to find out how youngsters learn and the best way for them to be taught, and we need enough money to change the physical environment we teach in, since most of our schools are like prisons. Also, we must come up with enough money to spend time to retrain everybody who presumes to teach now, when we find the best way to do it. Until we do these things I really don't think we are going to improve instruction, Stull Act or no Stull Act.

STATISTICAL AND MEASUREMENT PROBLEMS IN IMPLEMENTING THE STULL ACT

GENE V GLASS
University of Colorado

I shall discuss three means by which school districts might attempt to comply with the personnel evaluation section of the Stull Act. These three strategies were chosen for examination for quite different reasons--the first, standardized testing, because its use for teacher evaluation would be an egregious error; the second, controlled, simulated assessment of teachers' impact on pupil performance, because it has been recommended to the schoolmen of California; the third, observation and rating of teacher behavior and students' evaluation of teacher performance, because it is the method of choice under present circumstances.

Evaluating Teachers by Means of Pupil Gains on Standardized Tests

Although few evaluation experts would defend the practice, there is still danger that some school districts might use pupil gains on standardized achievement tests to evaluate teachers. The danger is sufficiently real that such attempts at evaluation deserve brief attention here. Evaluating teachers by measuring their pupils' gains from September to June on commercially available standardized tests is patently invalid and unfair. My faith is strong in the ability of standardized tests to uncover gross educational deficiencies in basic skills. But such instruments were not designed to reveal the variety of ways in which teaching and learning can be creative, favorably opportunistic, and uniquely meaningful to students. The inadequacy of standardized tests for the purpose of evaluating school learning has become a favorite theme of several contemporary educational researchers (Stake, 1971; Bormuth, 1970; Anderson, 1972). I accept their arguments against uses of standardized tests for purposes such as the one under consideration here. I urge those of you not familiar with the arguments to study them so that you can protect yourselves against thoughtless proposals to reward teachers for their students' grade-equivalent gains.

Aside from the irrelevance of much of the content of standardized achievement tests, their use in evaluating teachers is unjust. Nonrandomly constituted classes give teachers of brighter pupils an unfair advantage. This remains true whether the statistician calculates simple gains, residual gains, true gains, true residual gains, covariance adjustments, etc. Nothing short of random assignment of pupils to teachers as an ironclad administrative necessity would ensure that the teachers were in a fair race to produce pupil gains. Even random assignment of pupils to teachers is not a panacea. Chance would insure that in the long run a certain proportion of teachers would get an unfavorable draw--a Type I error, if you will. Type I errors seem relatively innocuous when a scientist runs one experiment, but they are frightening to contemplate when one realizes that there'll be 500 of them (250 of them making a good teacher look bad) when we evaluate 10,000 teachers ($\alpha = .05$). Random assignment of pupils to teachers begins to pall, and the scientific imagination wanders toward images of a standard class of 30 prepubescent robots in a mobile trailer moving from school to school--the perfectly controlled teacher evaluation device.

Three questions posed to me by the organizers of this conference bear on the point of separating teacher effects from extraneous effects, and they have a common answer:

1. How can student progress attributable to teachers, but not to other factors, be isolated and measured?
2. How can aspects of teachers that affect pupil progress be identified?
3. How can the effects on student progress of such factors as family educational background, family expectations and aspirations, and community factors be identified and measured? How can they be distinguished from the effects of teacher performance?

Attaining the three desiderata implicit in these questions is impossible, short of instituting enormously expensive simulation techniques. Either one randomly assigns pupils to teachers or he forfeits claims to having a fair evaluation system. No amount of post hoc statistical correction is equal to the demands for accuracy on a teacher evaluation system that has consequences as broad and serious as the system we are considering here. The first instance of a teacher losing pay or his job because his class mean fell three standard errors of estimate below a regression line is likely to lead straight to a lawsuit. Then we can count on experts in statistics and measurement giving conflicting testimony. The plaintiff's experts will maintain that the method by which the teacher was evaluated did not control for all preexisting inequalities among pupil groups; the defense experts will claim that it did. The judge will become impatient; the public will become cynical; psychometricians will blush.

I could pursue the answers to the problem of assessing the true teacher effects through complex, replicated experimental designs, through worries over distinctions between pupils and classes as the experimental unit, through concerns about order effects of repeated observations of teachers, through the specification of decision studies to insure adequate generalizability of the measurements, etc.--but I won't. It would be pointless of me to linger in Utopia when the immediate problem is upon us of finding a feasible, fair evaluation system that will work this year.

Even if the validity and fairness objections to using standardized tests could be met, available evidence indicates that teachers' effects on pupils' knowledge are not reliably measured by such tests. Assume that the teacher's true ability to teach traditional subject matter is a stable trait which fluctuates little from year to year. We shall measure this ability by administering a standardized test to the teacher's pupils and correcting for the initial status of the class by partialing out that portion of the end-of-year mean score which is linearly predictable from the beginning-of-year mean score (a procedure which has everything in its favor except logical good sense; see Cronbach and Furby, 1970). How stable from one year to the next will this "residualized" measure of the teacher's effect be? How subject is it to the vagaries of the shifting composition of the class across years, the fallibility of the test scores on individual pupils, and the compounding of measurement errors in the calculation of "gain scores?" In Table 1, I have integrated the findings from three unpublished studies which provide some answers to these questions. In each study the investigator calculated residual gain scores across September to May for two successive years of each teacher's class. The correlations of these gains across two years are "stability/reliability coefficients." They tell us how reliably we can thus measure what we assume to be teachers' true ability to teach their pupils what standardized tests measure.

If we were tempted to evaluate teachers with standardized tests before, we're not after seeing the reliability of such measures.

Table I

Summary of One-Year Stability Coefficients of Teacher Effects (Residual Mean Gain-Scores) on Standardized Test Performance

Nature of Subtest	Average r for the row	BENNET (1971) Grade Level: 1 Test: Metro. Ach. Covariate: Metro. Readiness No. of Teachers: n=34; r's	HARRIS (1968) Grade Level: 1 Test: Stanford Ach. Covariate: Stanford Achievement No. of Teachers: n=30; r's	HARRIS (1968) Grade Level: 2 Test: Metro. Ach. Covariate: Metro. Achievement No. of Teachers: n=24; r's	BROPHY (1972) Grade Level: 2 Test: Metro. Ach. Title I Schools No. of Teachers:		BROPHY (1972) Grade Level: 2 Test: Metro. Ach. Non-Title I Schools No. of Teachers:		BROPHY (1972) Grade Level: 3 Test: Metro. Ach. Title I Schools No. of Teachers:		BROPHY (1972) Grade Level: 3 Test: Metro. Ach. Non-Title I Schools No. of Teachers:	
					n=26	n=22	n=42	n=36	n=24	n=20	n=44	n=42
Reading	.23	.05^a	.32	-.08	.31	.00	.40^b	.42^b	.54^b	.39	.26	-.07
Arithmetic	.30	-.01^a			.24	-.12	.35^c	.39^c	.34^c	.42^c	.51^c	.56^c
Spelling	.49	.73^b	.52^b	.21								
Word Knowledge	.38	.23	.53^b (Word reading) -.01 (Vocabulary)	.19	.49^b	.18	.41^b	.42^b	.52^b	.78^b	.39^b	.45^b
Word Discrimination	.36	.69^b		.26	.26	.28	.63^b	.42^b	.28	.19	.30	.26

---

[a] An "outlying" case more than four standard deviations below the mean was removed before calculation of the correlation.

[b] Significantly different from zero at the .05 level (two-tailed test).

[c] Average of stability coefficients for arithmetic computation and arithmetic reasoning.

If the reader is still unconvinced of the inadvisability of attempting to evaluate teachers by means of pupil gains on standardized tests, he need only inspect the predictable results of such a policy. Teachers would teach the "safe" topics, possibly at the expense of the elusive but important ones. Teachers would not be permitted to administer their own standardized tests; an expensive external proctoring system would be required, and its costs in terms of trust might even exceed its costs in dollars.

We can anticipate confidently that evaluating teachers on pupils' gains on standardized tests would allocate rewards for "good teaching"--in fact, the superior gains would probably be due to superior pupil intelligence--in a malicious manner. Specifically, the large gains would occur in white, middle-class suburban classrooms, the areas least in need of competent teachers with high morale. The pattern of gain which could be expected can be seen in Table 2, in which are reported average residual gain scores for 1,245 fourth-grade pupils on the Metropolitan Achievement Test-Reading for various types of pupil categorizations (data from the 1969 Survey of Compensatory Education; Glass et al., 1970).

Teachers do not need additional incentives--which they would have if they were evaluated by their pupils' standardized test gains--to flee the tough teaching problems for the safe comfort of white, middle-class schools.

The PM^2 Method

W. James Popham and John D. McNeil pioneered in the study of determining teacher effectiveness by means of direct measurement of the teacher's impact on pupils' behavior

Table 2

1969 Title I Data; Grade 4; Metropolitan Achievement Test
N = 1,245

$$(g = y_{post} - \hat{\beta}_1 y_{pre} - \hat{\beta}_2 \text{ (elapsed time) } - \hat{\beta}_o)^*$$

	$\bar{g}$ (years)	n
Pupils:		
With pre-kindergarten experience	-.10	887
With no pre-kindergarten experience	.02	234
Family Annual Income:		
Under $3,000	-.09	250
$3,000-$4,500	-.09	335
$4,500-$6,000	-.03	513
$6,000-$7,500	-.02	155
$7,500-$9,000	.13	187
Over $9,000	.34	123
English Language:		
Not spoken in home	-.07	265
Spoken in home	.01	1328
Ethnic Group of Pupils:		
Black	-.11	753
Spanish	-.11	249
White	.18	594
Urbanism of Pupils' Home:		
Suburban	.02	1013
Urban and Suburban	.03	552
Rural	-.01	44
Occupational Level of Head of Household:		
Unskilled	.01	474
Semiskilled	-.04	405
Skilled	-.03	388
Owner-Manager	.07	173
Technical or Professional	.22	52

---

[*] In this equation,

g = residual gain score

y_{post} = end-of-year achievement

$\hat{\beta}$ = regression coefficients

y_{pre} = beginning-of-year achievement (knowledges, skills, and attitudes). Recently they have been joined in these endeavors by Jason Millman, who has teamed with Popham to offer school districts in California and across the nation such techniques for evaluating teachers. (For convenience, the Popham-McNeil-Millman method of assessing teacher effectiveness will be referred to in this paper as the P-M-M or PM^2 method.)

With the PM^2 method, a group of teachers is given advance notice (a few hours or a day or more) that they are to teach a particular topic to a group of pupils unfamiliar to the teacher. The teachers are often prepared by being allowed to study reference materials, instructional objectives, and even test items comparable to those which will be used as posttests. Frequently, the topic to be taught is one with which the teachers have had little prior experience. Ad hoc groups of between a half-dozen and 30 or more pupils are randomly formed from a large pool of pupils available for the assessment. The pupil group is taught for a period ranging between 30 minutes and an hour. Because of the random assignment, all teachers begin the instructional period with pupil groups randomly equated with respect to knowledge of the topic, ability, motivation, etc. Hence, posttest pupil performance is an unconfounded measure of the teacher effect; "simple gain," "residual gain," "percent gain," and similar scores are unnecessary. Teacher effects can be further studied by randomly reconstituting the ad hoc pupil groups, by changing the topic of instruction, etc.

Logical Arguments Advanced for the PM^2 Method

This basic technique underlies the system of measuring and evaluating teachers' effectiveness which Popham has proposed as a possible adjunct to the traditional methods for compliance with the teacher evaluation requirement of the

Stull Act. "Another possible addition to the Teacher Appraisal System, and one which many believe represents valuable augmentation of our appraisal techniques, is the use of a teaching performance test or, as it is sometimes called, an instructional minilesson," wrote Popham (1971a, p. 36) in Designing Teacher Evaluation Systems: A Series of Suggestions for Establishing Teacher Assessment Procedures as Required by the Stull Bill. Millman (1971, p. 15) indicated that his and Popham's company, Instructional Appraisal Services, began offering schools teacher performance testing for screening job applicants in the spring of 1972.

The major logical arguments advanced by Popham and Millman for the advantages of their conception of teacher evaluation are (1) that the PM^2 method of appraising teachers is "objective," whereas other methods (notably rating scales and observation schedules) are "subjective," and (2) that changes in pupil behavior like those which teachers produce during the simulated instructional periods are "the real thing" about schooling.*

The hobgoblin of "subjectivity" has been a dead issue in the philosophy of science for decades. "All measurement yields, not a property, intrinsic to the object being measured taken in isolation, but a relation between that object and the others serving as standards of measurement. When the relation is to other human beings, or even to the observer himself, it is not therefore a subjective one. As always, everything hinges on the controls which can be instituted, and on the sensitivity and reliability with which the discriminating judgments are being made" (Kaplan, 1964, p. 212). In scientific inquiry, every act of measurement, experimental design, hypothesis formulation, data interpretation, and so forth, involves human judgment; indeed these acts may involve aesthetic and, particularly, value judgments. In this sense, all science is subjective. The only "subjectivity" one need fear--both in science and in educational evaluation--is capricious and unsubstantiated opining. In this pejorative sense, the PM^2 method (in its selection of instructional topics, statement of objectives, and selection of test items) could be used as "subjectively" as the worst rating scales and observation schedules.

---

*For the complete arguments in support of the logic of the PM^2 method, see Popham's Designing Teacher Evaluation Systems (published by the Instructional Objectives Exchange, PO Box 24095, Los Angeles, Calif. 90024), which was written in response to the Stull Bill, and Millman's "Performance-based indicators of teaching effectiveness: Has the time come?" (an invited address at the annual meeting of the Northeastern Educational Research Association, November 12, 1971).

Second, is the teacher's ability to produce behavioral changes in pupils "the real thing" about schooling? Hardly. Education has many "real things." Changing children's behavior is just one of them.* Permitting children to grow in ways in which children have always grown when given supportive and interesting environments is another. Simple custodial care is a third "real thing" about schools that sounds trivial until one begins deducing the economic and social consequences of deschooling society.

Undeniably, however, changing pupils' behavior is one responsibility (among many) of teachers. However, we can currently conceive of and recognize in particular instances far more types of behavioral change than modern techniques can measure reliably. In spite of commendable success in measuring even complex cognitive processes, measurement specialists have ample cause for humility. Practical measurement of many affective and personality characteristics is mere hope than reality. Measurement of an educational goal as basic as creative (or at the very least "original") writing skill is still enormously complex and expensive. Teachers affect children's assessments of personal worth, which are closely linked to the children's feeling of satisfaction and dissatisfaction with themselves. Yet, few children trust so fully the adults they deal with in school that they will disclose their true feelings about themselves. Children learn something about what adults are from living with teachers, but whose faith is so strong that he believes the child's conception of the adult role can be elicited with a paper-and-pencil inventory?

---

*Popham (1971a, p. 39) quite appropriately acknowledged not only the multiplicity of educational outcomes but the need for multiple means of assessing teacher performance as well: "It should be emphasized, however, that the ability to accomplish pre-specified objectives in learners under short instructional periods represents only one criterion by which a teacher should be judged. The use of teacher performance tests simply reflects an additional criterion which might be employed in a local district's evaluation system."

Vignette

Teacher: "Today you are to answer some questions about what adults are like. Take one sheet off the top of the pile and pass the rest back. Be certain that your name appears in the upper right-hand corner. Use only a No. 2 pencil and mark heavily between the lines. There'll be no conversation during the test. You'll have 30 minutes to answer the questions on the sheet.

Are there any questions? Good. You may begin."

Pupil (to himself): "Obviously adults are rigid and demanding; they hide behind regulations, and they're typically suspicious. They aren't sincere when they ask for frank opinions.

Now let's see; what do they want us to say on this thing?"

Millman (1971, p.10) asked, " . . . do we have any assurance that the teacher who can't communicate information effectively with her pupils will be able to assist them, for example, 'in their pursuit for personhood'? I rather doubt it." So do I; but I doubt the converse more readily than Millman does, I suspect. The "effective information communicator" may be no better as an adult role model or steward of authentic children than the less effective "information communicator."

As Paul Meehl showed long ago, there comes a point in the error rates of psychological measures where the use of such measures may do more harm than good. For example, a personality inventory that incorrectly labels ten merely antisocial individuals as dangerously paranoid schizophrenic for every one dangerous paranoid schizophrenic it correctly identifies is probably worse than no personality inventory at all. A blanket diagnosis of no one as dangerously paranoid schizophrenic may be more acceptable, considering the injustice of institutionalizing people who are "different" rather than sick.

Moral: Educational evaluators must know when to pick up their tools, and when to lay them down.

Empirical Evaluation of the PM^2 Method

The testimonies for the empirical validity of the PM^2 method have been imprudent and immodest. Justiz (1969, p. viii) claimed for his study that it had produced "the first reliable measure of general teaching ability." Indeed, he thought so highly of the technique that he proposed a handy, eight-step adaptation of the method for use in evaluating junior college faculty. Connor (1969, p.16) wrote that his study not only "validated" the teacher performance assessment technique but also verified "the construct of instructional proficiency as a measurable variable." Popham (1971b, p.116) cited Justiz and Connor as having "recently reported high positive correlations between teachers' achievements on two different short-term performance tests. . . . Such results are, of course, encouraging." McNeil (1972, p.622) considered such performance tests to be an "answer to the problem of identifying the effective instructor."

In the face of this partisan enthusiasm for the PM^2 technique, I wish to examine the radical antithesis, namely that the technique has not been shown to possess nonzero reliability when used as advocated by Popham and others as a means of satisfying the Stull Act requirements. In the following pages I will attempt to demonstrate that the very data which are said to establish the reliability of the PM^2 method are consistent with the claim that the method has zero reliability for generalizing across factors for which it must show stable measurements. A careful examination of the data obtained by Connor, Justiz, Popham, McNeil, and others will reveal no evidence to disprove the assertion that the teacher's effect on pupil performance is measured with near-zero reliability across both topics taught and differing groups of pupils by the PM^2 techniques. In short, I shall argue that the PM^2 technique has yet to pass the minimum requirement for measurement utility.

To be useful in implementing the Stull Act teacher evaluation requirement, the PM^2 method must show substantial reliability across different topics and different groups of pupils. The construct "good teaching" forces the issue of "topics" generalizability. "Good teaching" means more than the ability to teach isolated curiosities of the curriculum. If "good teaching" is not evident in the same teacher's consistent impact across a reasonably broad range of topics (even within a domain itself as narrow as secondary school social studies, for example), then "good teaching" is not actually being measured. The PM^2 method must yield stable measures across different (possibly randomly constituted) groups of pupils for practical reasons. If teachers produce highly unstable effects when teaching the same topic to several different pupil groups, fair comparisons among teachers can only be bought under two conditions (the first, fatuous; and the second, cumbersome and impractical): (1) all teachers in a district about whom decisions are to be made must teach the same standard groups of pupils; or (2) each teacher's effect must be measured by averaging his or her efforts across many randomly different groups of pupils. The PM^2 method either shows reliability across topics and pupil groups or else it remains a laboratory technique, useful for teacher training and research, perhaps, but not up to the standards of utility, economy, and fidelity required for the purpose we are considering.

A detailed analysis of the pertinent studies follows.

The Rosenshine Review

Rosenshine (1970) compiled the results of several studies of the short-term stability of the teacher effect that used methods like those of the PM^2 technique. The relevant results are reproduced in Table 3.

I have not carefully examined the studies by J. C. Fortune (Fortune, 1966, 1967; Fortune et al., 1966). However, later in this paper the Justiz study is examined in detail; I think its findings should be discounted. The Fortune studies reveal a surprising variability in stability coefficients. The reliabilities range from -.45 to .82 and average .13.

In citing Rosenshine's review, Millman (1972, p.12)

Table 3

Correlations of Teacher Effects across Two Topics and Two Different Pupil Groups (after Rosenshine, 1970)

Study	Pupil Grade Level	Instructional Topics	No. of Teachers	Stability Coefficients $\underline{r}$		
Fortune (1966)	Head Start	Social Studies	15	.21	.42	.45
	Head Start	Social Studies	15	-.08	.13	.25
Fortune (1967)	4-6	Social Studies	13	-.45	-.42	.42
	4-6	English	15	-.35	-.07	.53
	4-6	Math	14	-.29	.05	.82[a]
Fortune et al. (1966)	7-9	Social Studies	40	-.10	.01	.43
Justiz (1969)	10-12	News Story Structure and Computer Punch Card	9		.58[a]	
Justiz (1969)	10-12	News Story Structure and Computer Punch Card	5		.87	

---

[a] Significantly nonzero at the .05 level, two-tailed test. All other coefficients reported in this table are nonsignificant at the .05 level.

wrote:

"The best summary of the best of the research literature . . . has been done by Rosenshine (1970). There is some consistency in teaching competency as measured by pupil growth, even when the teacher teaches different topics to different students. In the most controlled studies, correlations around .40 to .50 are often found." [Italics added.] (Millman neglected to mention, however, that correlations around -.40 are also often found.) Millman did not point out that only 3 of the 20 coefficients reported by Rosenshine were statistically significant at the .05 level.

The Justiz Studies

Justiz (1969) conducted two studies in two separate high schools. In School A, ten student-teachers each taught a group of about 20 pupils a 30-minute lesson on "new story structure"; the pupils were drawn almost entirely at random from a pool of available pupils in the school (the only compromise being that no teacher was assigned one of his own pupils). Pupils were tested over the content of instruction for 15 minutes after the lesson. The pupil groups were reconstituted (again nearly at random), and the student teachers instructed their new pupils on "concepts of the computer punch card." A 15-minute posttest was given, as before. The study was completed within two hours.

Justiz's scoring of the posttests is not clear; apparently the "new story structure" test involved rearrangement or ranking items which were scored by accumulating the sum of squared discrepancies between each pupil's responses and the keyed responses; the "punch-card" test appears to have been scored by "number correct." The pupil-group means (possibly on a percentage scale) were reported for nine of the ten student teachers.

Student Teacher	"News Story" Mean (X)	Rank on X	"Punch-Card" Mean (Y)	Rank on Y
a	81.5	1	64.5	3
b	71.5	2	70.5	2
c	67.5	3	55.0	6
d	66.0	4	82.0	1
e	61.0	5	43.0	8
f	57.5	6	56.0	4
g	57.3	7	44.0	7
h	57.2	8	55.5	5
i	57.2	9	30.0	9

$\underline{r}_{xy} = .58$; $p < .05$; 95% confidence interval (.08, .89)

Justiz (1969) reported a Spearman rank-order correlation of .64 for these data, which I have confirmed. A Pearson's $\underline{r}$ seems equally, or more, appropriate; its value for the data is 0.5, which is just barely statistically significant ($\alpha = .05$, two-tailed test). But the 95% confidence interval for $r = 0.58$ with $\underline{n} = 9$ better reflects how accurately the teacher-effect stability coefficient has been determined in this study--the interval extends from .08 to .89, a broad range indeed. Thus we learn that the teacher effect is somewhere between almost completely unstable and perfectly stable across two topics.

The influence of one complication in Justiz's study is difficult to assess. One of the ten student teachers was dropped before the end of the study. The reason given was that this student teacher " . . . refused to teach the Punched-Card Computer Concepts subject for lack of preparation" (Justiz, 1969, p. 47). Presumably the teacher was prepared and did teach the "news story" lesson. If the teacher had been required to participate in the entire study--as might be expected in a system of mandatory teacher evaluation with this method--then we could expect some discrepant performance of his pupils between the "news story" lesson, for which he felt prepared, and the "punch-card" lesson, which he refused to teach. To illustrate the potential influence on the stability coefficient of having dropped this teacher, assume that his pupils would have achieved an average score on the "news story" lesson and the poorest score on the "punch-card" lesson. The rankings would have been changed as follows:

Student Teacher	Rank on X	Rank on Y
a	1	3
b	2	2
c	3	6
d	4	1
"j"	5	10
e	6	8
f	7	4
g	8	7
h	9	5
i	10	9

$r_s = .47$; $p > .10$

The value of Spearman's rank-order correlation coefficient on the above data is 0.47, which is not significant even for $\alpha = .10$, two-tailed test. Obviously, one missing case can have a marked effect in such a small sample.

In a second high school, School B, Justiz repeated essentially the same procedures with the same topics, one exception being that teachers instructed the same group of pupils on the two topics. The sample included seven student-teachers. Two of these teachers were dropped before completion of the study because of irregularities in posttesting.* The posttest data for the five student teachers' "classes" were as follows:

---

*Rosenshine's (1970) reported sample sizes of $n = 10$ and $n = 7$ for Justiz's study; in fact, they were $n = 9$ and $n = 5$. The confusion is due to the fact that Justiz's reports of sample sizes varied from one publication to the other.

Student Teacher	"News Story" Mean (X)	Rank on X	"Punch Card" Mean (Y)	Rank on Y
a	70.0	1	57.5	2
b	66.0	2	71.5	1
c	59.5	3	51.5	3
d	56.7	4	47.5	4 or 5 or 4.5
e	51.5	5	47.5	5 or 4 or 4.5

$r_s = .90$, $p = .10$ or $r_s = 80$, $p > .10$ or $r_s = .87$, $p > .10$.

Justiz reported a Spearman rank-order correlation of .90, "significant at the .05 level of confidence." The calculations and conclusion deserve closer scrutiny. For $n = 5$, an $r_s = .90$ is significant at only the $\alpha = .10$ level for a two-tailed test; the two-tailed test is justified in this instance in view of Fortune's frequent finding of a negative stability coefficient in similar studies (see Rosenshine, 1970).

But there's a more fundamental problem here (if the reader is unimpressed with quibbles over .05 versus .10 levels of significance). Justiz broke a tie on variable Y between teachers d and e in favor of a higher correlation. If the tie is broken in the other direction, the value of r_s is .80, which with $n = 5$ is nonsignificant at any reasonable level of α. The best resolution of the tie is to assign a rank of 4.5 to both student teachers, which yields a third nonsignificant r_s of 0.87.

Conclusion: disappearing subjects, miniscule sample sizes, errant correlation coefficients, and nonsignificant statistics do not yield the "amazing consistency and statistical significance" that Rosenshine (1970, p. 659) saw in the Justiz studies.

The McNeil Study

McNeil (1972) reported on data collected at a "teaching fair" which bear on the question of PM^2 method reliability. Groups of teachers were randomly paired with groups of three elementary school pupils. A topic (e.g., code breaking) was taught, the pupils were excused for a short recess, and a second group of three pupils was randomly constituted for the second instructional period. The instructional topic may not have changed between the two instructional periods.

McNeil's (1972, p. 626) results section was brief and is reproduced in toto here:

> "First session scores of fifteen teachers on Test I were compared with these teachers' scores from a second session. The correlation between the scores was .521, significant at the .05 level. Also, correlations between teachers' performances on the different tests given ten weeks apart with different kinds of pupils were positive. For instance, thirty teachers took Test 3 at the end of their methods course and completed Test 4 ten weeks later after a student teaching assignment. Their scores showed a Pearsonian r of .388 (p < .05). As indicated in Table I . . . one could have made a probable prediction about the likelihood of high achieving and low achieving teachers (top 25 percent and bottom 25 percent) making a similar showing on a second test weeks later. The chances of these teachers maintaining their level were greater than three to one."

TABLE I

Chi Squares for High and Low Teacher Performance on Tests of Ability to Teach Two Different Tasks of Reading

		TASK(S) 4		
		High	Low	
Task 3	High	9	4	χ^2 = 4.54
	Low	4	9	
		High	Low	
Task 2	High	12	4	χ^2 = 4.40
	Low	6	10	

McNeil's report of results is somewhat fragmentary. It appears that the correlation coefficient, .52, reported in the second sentence of the results is a stability coefficient for the same instructional topic across two sessions. As such, it is comparable to the coefficients Rosenshine (1970) reported for fixed topic reliability studies. The 95% confidence interval around an r of .52 (n = 15) extends from just above 0 to .80, a broad range indeed. However, the critical question is one of stability of the teacher effect across topics. McNeil refers to the correlations of this type which he has computed as being "positive." The one instance which he reports in detail (between Tests 3 and 4 with ten weeks intervening) produced an r of .39 (n = 30): the 95% confidence interval is .03 to .62. It is crucial to know whether the one instance reported was the most favorable finding among the "positive" correlations McNeil calculated, or whether it is genuinely representative of all of the coefficients, but we do not have this information.

McNeil's Table I first appears to be based on the r of .39 to which he referred in the text of his paper. However, the situation is unclear, principally because the original sample contained 30 teachers and the table, which is said to contain half of the teachers ("top 25 percent and bottom 25 percent"), is based on 26 and 32 teachers instead of 15. We must presume that the table contains data from studies independent of the study which yielded the r of .39.

Both contingency tables in McNeil's Table I are said to yield a statistically significant association (χ^2 test, α = .05) between high and low teacher effectiveness across the two topics. My calculations show otherwise, however. McNeil reported a χ^2 of 4.54 for the Task 3-Task 4 comparison. My calculations yield a χ_1^2 of 3.85 without the continuity correction--a figure which should have responded exactly to McNeil's--and a χ_1^2 of 2.46 with correction. The continuity correction is clearly in order due to the small sample size; the χ_1^2 of 2.46 is not significant even at $\alpha = 10$. (Fisher's exact test similarly failed to reject with $\alpha = .10$).

For the lower half of McNeil's Table I, the value of χ_1^2 corrected for discontinuity is 3.17, which is nonsignificant with $\alpha = .05$. (However, Fisher's exact test yields a probability of .04 of a more extreme occurrence under the null hypothesis of no association.)

Fortunately it is possible to approximate the Pearson r's for the data upon which McNeil's Table I is based. His dichotomizing of the variables into upper and lower quarters is close to the "upper-lower 27%" splits employed years ago in item analysis. Tables and nomographs exist (see Colver, 1959) for converting fourfold continency tables based on the upper and lower 27 percents of each distribution into approximate Pearson r's. The values of r for McNeil's data are .40 and .38 for the top and bottom portions of Table I respectively. These values are statistically nonsignificant at all reasonable α-levels (see Ross and Weitzman, 1964).

In short, the status and full import of McNeil's data are unclear. The confidence intervals are broad, the nonsignificant findings also need to be reported, and some of the results reported as statistically stable appear not to be.

The Connor Study

Connor (1969) arranged for 17 experienced male teachers to instruct subgroups of 77 high-school pupils in the topics "propaganda techniques" and "chance (probability)." The teachers were not particularly familiar with the topics and had very little time to prepare their lessons. A systematic, but probably nearly random, method of assigning five pupils to each teacher was employed, with the condition that no teacher was assigned a pupil with whom he was acquainted.

Pupils in each "class" were tested on "propaganda techniques" prior to a 50-minute instructional period. A comparable posttest followed instruction. Later in the day, the pupils were reassigned to teachers and the new ad hoc classes were taught "chance" for 50 minutes; pretests and posttests were again administered.

A pre- and posttest gain score was calculated by subtracting the pretest number of items correct (times the number of pupils) from the posttest number of items correct (times the number of pupils) and dividing by the difference between the highest possible score and the pretest score. Thus, "gain as a percent of possible gain" was used as the measure of each teacher's effect on pupil performance. Use of the pretest in measuring the teacher effect was unnecessary because of random assignment of pupils to groups, and it probably lowered the reliability of the measurements to some unknown extent.

Connor (1969, p. 14) presented the following raw data:

Teacher	Propaganda (X)	Rank on X	Chance (Y)	Rank on Y[a]
1	.3636	4	.2857	5
2	-.0833	16	.5000	1
3	.1111	12	.2500	6
4	.0000	15	.0000	15
5	.1818	9	.1538	11
6	.0556	14	.1538	12
7	.2143	8	.2000	7
8	.2308	7	.1818	8
9	.2667	6	.3077	3
10	.6667	2	.1429	13
11	.3333	5	.1667	10
12	.6363	3	.2857	4
13	-.1538	17	.0909	14
14	.0769	13	.1667	9
15	.1538	11	-.2222	17
16	.7692	1	.3846	2
17	.1667	10	.0000	16

---

[a] Ties randomly broken.

Connor did not report the value of either r or r_s, i.e., Spearman's rank-order coefficient. They are small and insignificant, however.

Clearly, Connor's data give no evidence of nonzero stability across topics. Instead of reporting either of the above two coefficients, Connor reported a coefficient "corrected for attenuation" of $r'_{xy} = .72$. He applied an inappropriate significance test to this value, testing it as one would test a typical Pearson's r on 17 cases. Appropriate inferential tests (see Rogers, 1971, or Forsyth and Feldt, 1969) reveal that even Connor's corrected coefficient is statistically nonsignificant at any reasonable α-level.

Moreover, the correction for attenuation procedure was inappropriate. Test reliabilities were estimated from internal consistency type coefficients and do not embody a meaningful definition of measurement error for the purposes of the correction. More importantly, however, correlations corrected for attenuation may have conceptual meaning but are seldom practically useful. There's small comfort in the fact that true scores on X and Y correlate highly when we're being asked to make decisions on the only available data, viz., the fallible observed scores.

The Belgard, Rosenshine, and Gage Study

Forty-three teachers taught lessons on the economic, political, and social conditions of Yugoslavia and Thailand to their regular classes of students. Class sizes varied from 10 to 31 pupils. One week before the two instructional periods, each teacher received copies of the instructional materials, together with half the items from the posttests to be administered to pupils after instruction. The lessons were taught for two 15-minute sessions on successive days; pupils were tested after each instructional period.

In an attempt to equate the abilities of the nonrandomly constituted pupil groups, Belgard, Rosenshine, and Gage (1971) administered a tape-recorded lesson on Israel to all classes on the third day of the study. Pupils' performance on a subsequent posttest was then used as a controlling variable for equating the 43 classes. Essentially, residual "gain scores" --the difference between the class mean on the Yugoslavia test, for example, and the regression estimate of this class mean from the Israel class mean--were computed and correlated across the two topics to measure the reliability of the teacher effect.*

At this point, two fundamental questions can be raised. The study did not involve random assignment of pupils to teachers; instead, an attempt was made to control for pupil differences with ex post facto statistical controls. One cannot be certain that the equivalence sought was actually achieved. It is likely that the classes still differed to a nonrandom degree on variables not controlled statistically (e.g., some specific aspects of learning ability or motivation). On the other hand, observation of the control variable (i.e., performance on the Israel test) was made after the two teacher-directed instructional periods. The possibility exists that the statistical equating actually removed stable variance in the teacher effect present in the variance of the Israel class means because of positive transfer effects, motivation carry-over effects, etc., from the two previous lessons. It would have been preferable to observe the control variable (through testing of pupil learning about Israel from the standard tape-recorded lesson) *prior* to the teacher-directed instructional periods. The first influence (lack of complete equating of classes) may have spuriously inflated Belgard, Rosenshine, and Gage's stability coefficient; the latter influence may have deflated it. We cannot be certain which influence was stronger.

---

*In fact, the investigators calculated the partial correlation between the class means on the Yugoslavia and Thailand tests with the Israel class mean held constant. This is equivalent to correlating the residuals of the Yugoslavia and Thailand variables formed by removing variance due to the Israel means.

Belgard, Rosenshine, and Gage reported a correlation of the Yugoslavia and Thailand residual class means of .47 (t = 3.37 with df = 40; $p < .01$). The investigators recognized that this coefficient is based on different topics taught to the *same* pupils--not a practical condition for implementing the Stull Act, incidentally. To correct for this condition, they divided the data in each class in half, designating even- and odd-numbered pupils. The residual scores were then recalculated separately for the two halves of each class and correlated across topics. The correlation of Thailand residual means for the even-numbered pupils with Yugoslavia residual means for the odd-numbered pupils (or vice versa) is then a measure of teacher stability across both topics and pupil groups. The authors reported the following two stability coefficients:

> Correlation of Yugoslavia residual means for odd-numbered pupils with Thailand residual means for even-numbered pupils = .16
>
> (t = 2.58, df = 40; $p < .01$)
>
> Correlation of Thailand residual means for odd-numbered pupils with Yugoslavia residual means for even-numbered pupils ($p < .01$) = .38

Belgard, Rosenshine, and Gage then applied the Spearman-Brown formula to estimate the size of these two coefficients for classes twice as large as the "half-classes" in which they were derived; this technique is perfectly appropriate. The Spearman-Brown "stepped-up" stability coefficients equaled .28 and .55. They averaged the two figures and settled on a stability coefficient of .41. But this is one of those interesting cases in which a set of scores (two, in this instance) does not "tend centrally" very well, and the mean value obscures more than it illuminates. The two methods of estimating the stability coefficient produced an extremely large discrepancy, though they were estimating the same parameter value.

Belgard's, Rosenshine's, and Gage's (1971, p. 185) conclusion seems most appropriate:

> "The correlations were not high enough to indicate that the effectiveness of individual teachers can be measured with adequate reliability with only two lessons, ten-item tests, and classes of about 21 students. For such reliability, higher than about .40, additional lessons, longer tests, and larger classes would be needed."

The Popham Studies

Although Popham's (1971b) comparisons of experienced and inexperienced teachers or experienced teachers and laymen with respect to their relative impacts on pupil performance were not addressed directly to the question of the reliability of the measurement techniques, they provide collateral evidence on the question.

Popham arranged for groups of experienced teachers, college students, and laymen to teach various topics to groups of high-school and junior-college students. A variation of "randomized blocking" insured equivalence of the student groups taught by the teacher and nonteacher groups. Popham presented three comparisons of teacher and nonteacher performance in terms of the average "percent items correct" earned by the students:

Topic	Comparison Groups	Number of "Instructors"	Average Percent Items Correct
Social Science (research methods)	Experienced teachers	13	67%
	College students	13	65
Auto mechanics (carburetion)	Experienced teachers	28	49
	Garage mechanics	28	47
Electronics (basic power supplies)	Experienced teachers	16	52
	TV repairmen and electronics workers	16	50

Popham (1971, p. 601) wrote that: "Although space limitations preclude an examination of possible methodological defects which might contribute to these results, there appear to be no readily available loopholes by which we can explain away the nonsignificant outcomes. A more straightforward explanation is available. Experienced teachers are not particularly skilled at bringing about specified behavior changes in learners."

A second explanation is both consistent with the available data concerning the reliability of the PM^2 method and more generous in its implications for the teaching profession. Zero reliability measurement techniques will not show statistically significant discriminations among any groups; at least such differences will not be found at a rate greater than the size, α, of the significance test. Nonzero measurement reliability is a necessary condition for discriminating among groups on the measure. For example, a test of "heads flipping ability" in which one's score is the number of "heads" he tosses in 100 flips of a fair coin--a test with zero reliability in the stability sense--will not show consistent, significant differences among streetcar conductors, waitresses, bank cashiers, or "men on the street."

Popham's conclusion may be true. On the other hand, his measurement procedure may be nearly completely unreliable and incapable of revealing genuine differences among any groups.

Concluding Assessment of PM^2 Studies

The PM^2 group may have stacked the cards against showing reliability for their technique. In most cases, the teachers must teach novel material (e.g., the Hollerith computer punch card, or new story structure). Thus, no teacher "has an advantage" over other teachers due to familiarity with material. In fact, the use of novel material may take away from teachers the very "advantage" which makes one teacher better--and consistently better--than another. Teachers' knowledge of the material they teach is certainly a stable characteristic and may, particularly at higher grade levels, determine--in some part at least--their effectiveness. Perhaps PM^2 and others (Fortune; Belgard et al.) have systematically partialed out of their studies an important stable variable related to teacher effectiveness. On the other hand, a few studies (McNeil, 1972, for example) used material more familiar to the teachers (e.g., teaching multiplication of two-digit numbers or pronouncing initial vowels in words with a silent final "e") and failed to produce convincing evidence of stability of teacher effects; moreover, convincing evidence is lacking that depth of teachers' cognitive knowledge of a subject (beyond some minimal level) is related to their pupils' progress in knowledge acquisition.

But fine points about naive versus knowledgeable teachers cannot do much to brighten the dismal composite picture in Figure 1 of the results of the available short-term stability studies. Presented in Figure 1 are the 95% confidence intervals on the 21 coefficients which can be found in the published literature. McNeil's two nonsignificant coefficients are not included, since exact confidence intervals

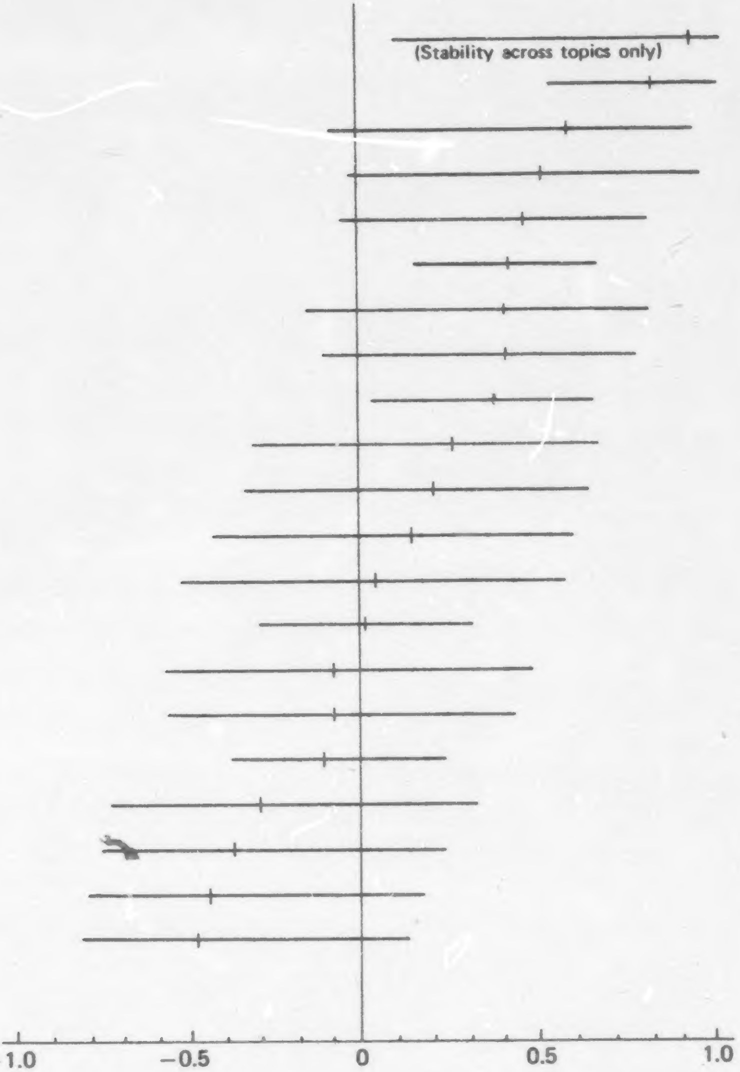

Figure 1. Summary of estimations (95% confidence intervals) of the stability of the teacher effect across instructional topics and pupil groups.

can't be drawn. Only four of the intervals fail to span zero. All but one span .15.

In theory, low reliability problems are among the simplest measurement problems to solve. Given that in its present form a measurement procedure has nonzero reliability, no matter how low its reliability may be it can be made more reliable by lengthening it, i.e., duplicating and averaging parallel measurement from the procedure. The Spearman-Brown formula allows us to calculate the reliability of a procedure lengthened k times from knowledge of its initial reliability, p'_{XX}.

Assume that from the discouraging picture of the reliability (across topics and pupil groups) of the PM^2 technique one draws the most generous conclusion consistent with good sense, to wit, that the technique has a reliability in the region $p_{XX}' = .20$ to .30. These estimates are based on the conditions that obtained in the studies reviewed above, the most important condition for our present purpose being that the entire procedure per teacher measurement involved at least one hour of time for about a dozen pupils, at least three or four hours time per teacher, and several hours of professional time for development and administration per episode. Figure 2 is a graph in which the reliability of the PM^2 technique is related to the lengthening of the method required for reaching various levels of reliability. For instance, if the PM^2 truly has initial reliability .30, it would have to be lengthened 10 times to achieve a composite reliability of .81. In other words, the simple methods typical of the studies reviewed above would have to be repeated ten times across ten different instructional topics before the average for a single teacher attained a reliability comparable to the one-year stability reliability of the typical standardized achievement test that we administer to our pupils.

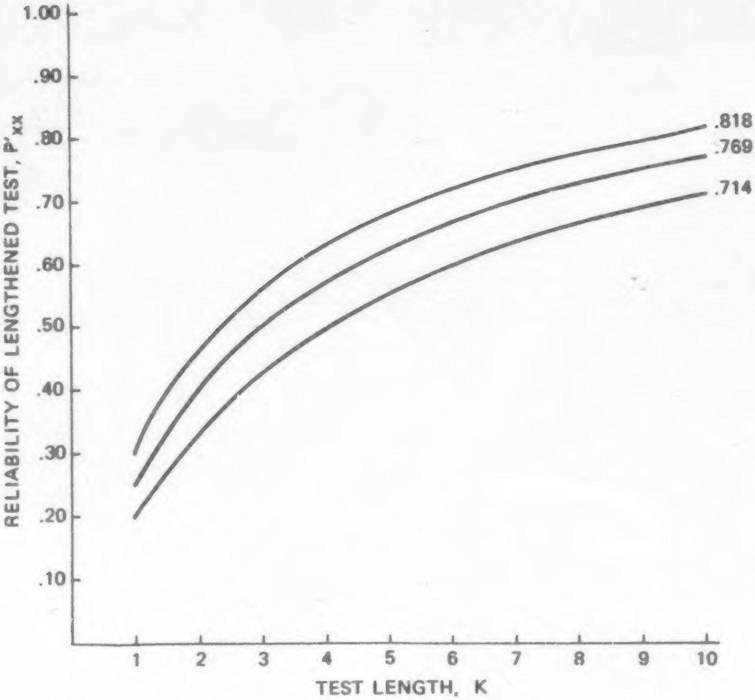

Figure 2. Relationship between test length and test reliability for tests with unit length reliabilities of .20, .25, and .30.

Net costs? Staggering, if we even bothered to pause and calculate them. Add to this gloomy picture the developmental costs of devising new topics and tests each year (obviously the topics are "consumable" year in and year out) and the administrative costs of shuffling pupils about (after all, the procedure is unfair unless the pupil groups are randomly equivalent across all teachers about whom differential decisions are to be made from the measurements), and we behold an evaluation system of enormous proportions.

There are few eternal verities in the corpus of educational research findings. Girls read better than boys;

intelligence correlates positively with achievement: "As it was in the beginning, is now and ever shall be, world without end. Amen."

I wish to nominate a third principle for promotion to the status of a great truth. A. S. Barr concluded in 1929 after a study of teaching behaviors that instability of teachers' behavior from one lesson to another is the dominant source of unreliability in the effect of teachers' actions on student learning. The contemporary studies reviewed above prove Barr's conclusion beyond a reasonable doubt.

No conclusion is more fitting for this section than a quotation of Popham's own prudent remarks:

> It now appears, in light of the grossness of the measurement devices likely to be available in the near future, that we shall be pleased even if the performance tests are suitable for use only with groups. In other words, it will be a sufficient advance to develop a reliable group criterion measure which could be used in myriad educational situations such as to assess the efficiency of teacher education programs. (Popham, 1971b, p. 116.) [Italics added.]

An Observational-Judgmental System

If not through standardized tests, if not by means of the PM^2 simulation method, then how should teachers be evaluated to comply with the Stull Act? Searching for an answer has been a personally significant experience. Initially I was temperamentally inclined toward the PM^2 arguments and methods, but the data convinced me that the methods should not now be adopted as the exclusive means of teacher evaluation. The alternative of evaluating teachers by means of observer ratings of teachers' classroom behavior and student feedback came to mind, but it first raised unpleasant associations with the pettiness and irrelevance of traditional "supervision" and accreditation. Nonetheless, more serious thought reveals this alternative to be the most feasible system of teacher evaluation with any claim to validity. Indeed it is the only system about which one can talk realistically of wide-scale adoption in California for several years to come.

In proposing teacher evaluation based on pupil and outside observer judgment, I am not recommending a ragtag system of uncritically collected data of dubious validity. The properties of the data should be studied periodically by means of the various statistical tools available for examining the reliability and validity of measurements. If two independent observers looking at adequate samples of teachers' performance can't agree on their judgments of teachers' "use of appropriate illustrations," then work needs to be done with the judges, the rating scales, the methods of sampling teacher behavior, or all three. Finding that principals' ratings of teacher rapport with pupils do not correlate with pupils' expressions of rapport with the teachers casts doubt on the principals' ratings, the pupils' ratings, or both--and something must be done about the situation. Would such observational data comply with the requirements of the Stull Act or must compliance entail--as some have implied--measurement of pupil performance and attribution of the just increment in the performance to the teacher responsible? Perhaps strict constructionist and liberal interpretations of the Stull Act will arise over the next several months; I, for one, construe the phraseology ". . . assessment of certificated personnel competence as it relates to . . . " standards of expected pupil progress to permit evaluations based solely on observation of teachers' classroom behavior.

The observational-judgmental system envisioned has three principal elements: (1) trained observers' ratings of teachers' relevant classroom behaviors; (2) students' evaluations of teachers; and (3) collateral data.

Observation of Teachers' Classroom Behavior

The word "relevant" distinguishes the first component of this judgmental system from our pathetic history of attempting to rate teachers' "openness," "professional manner," or "sensitivity." Past failures to rate teacher behavior reliably stemmed from vague, nonspecific definitions of the behaviors to be rated and lack of training of raters. When at least these two defects are remedied, ratings reach surprisingly high levels of consistency and long-term stability (see Medley and Mitzel, in Gage, 1963, p. 292, for example). I would propose that no characteristic of teaching be incorporated into the rating scales until research has established both that it can be reliably observed and that it bears some significant relationship to desired pupil cognitive and affective states. This requirement for research validation is not a purist's "cop-out." The ceremonial enunciation of high standards by researchers usually precedes requests for more time and money so that the researchers can establish a firm knowledge base for policy. Not so in this case. We are ready today to suggest the types of teacher behavior that research would support in a judgment system. In Designing Teacher Evaluation Systems, Popham (1971a, p. 9) wrote that "ratings generally do not correlate particularly well with measures of pupil growth." Quite the contrary; they not only correlate consistently with pupil "cognitive" growth--the special sense in which Popham's remarks must surely have been intended--but these correlations are higher than the correlation of the PM^2 technique with itself across even quite closely related topics! Research evidence (see Rosenshine, 1971) currently justifies observing and judging the following teacher behaviors: clarity of presentations and explanations, enthusiasm, variety in use of instructional materials and techniques, task orientation and "businesslike" behavior, and provision of ample learning opportunities. This same research suggests observation of many other potentially important teacher behaviors: teacher use of students' ideas, use of multiple levels of discourse, use of criticism, and "probing." Each characteristic can be given precise definition and can be reliably measured by trained observers. The privilege of elaborating this work and advising on its installation in the schools belongs to Gage, Flanders, Rosenshine, Nuthall, and other researchers who have pioneered in this territory and now see their efforts beginning to bear fruit.

Pupils' Evaluations of Teachers

The second element of the observational-judgmental system involves the collection of student evaluations of their teachers. The data collected from students could be of several types, probably including, as a minimum, information to corroborate outside observers' ratings of teacher behaviors. Secondly, the system could include students' judgments of the learning environment (as defined and studied by Walberg and Anderson, 1968, 1969).

Walberg and Anderson's work on the "Learning Environment Inventory" seems particularly promising and immediately useful. Their inventory, which students fill out, gives class mean scores on 14 factor analytically derived dimensions of learning environment, each score being derived from individual pupil responses to seven items. The names of the dimensions and their internal consistency reliabilities (calculated on class means and in a somewhat restricted sample, incidentally) invite immediate utilization and further research: (1) Intimacy (.78); (2) Friction (.78); (3) Satisfaction (.80); (4) Difficulty (.66); (5) Apathy (.83); (6) Favoritism (.77); (7) Goal Direction (.86); (8) Disorganization (.81); and so forth.

Third, pupils' contribution to the observational-judgmental system could include their reports on the state of basic human decency that exists in the classroom. I expect some of my colleagues to scorn such soft-headed suggestions as evaluating teachers on whether they are "decent human beings." McClellan (1971) reminded us that all teaching is simultaneously a "saying, doing and making." We must judge what teachers "do" to children as well as what they "make" of them. Even if the actions had no visible residual effects on the pupils' adult behavior, teachers' rudeness, bad manners, and ill-tempered repression would be contemptible in themselves. Bane and Jencks' (1972, p. 42) pessimism regarding the school's impact on children's ultimate stations in life led them to conclude that a school should be evaluated on whether "the students and teachers find it a satisfying place to be." Scriven (1972, pp. 215-216) put it better: "If school were a pill taken in an instant, it would be appropriate--morale consideration apart--to evaluate it on its effects alone. But schools go on for a long time by any standard, and almost interminably by youthful standards which are relevant in this case. Consequently one must recognize the necessity for regarding the consequences of the treatment as including the consequences of lengthy incarceration within the system.* This is often

---

[*] I would add here that the mode of incarceration must be considered even when it has no discernible "consequences." We would not hesitate to render a moral judgment of a corrupt legal system that sent an innocent man to prison for 13 years even if the imprisonment had no permanent effect on him.

a heavy cost as seen by the child, and I believe the most striking residue of paternalism in evaluation is the failure to include it in most studies. . . . The failure to include these criteria is also partly due to the effects of the doctrine that good medicine often tastes bitter. . . . If there is considerable doubt about the merits of the pill then there is something to be said for making it taste good, especially if you're forcing it down an unwilling throat."

Pupils themselves are the only reliable reporters of those internal mental states which, even if as evanescent as this morning's hangover, nonetheless count heavily in what makes school good or bad.

One must not forget that pupils are not the only human beings in the "schooling" system. By conservative estimate there must be over two million adults who, as teachers, spend more time in schools than pupils do. The philosophy that regards "changes in pupil behavior" as the "real thing" about education treats these two million adults as means, and tends to disregard them as ends in themselves. I would dwell on the need to protect the teachers, but they seem quite able to protect themselves these days.

Collateral Data

I have omitted from this teacher evaluation system any data of the type upon which teaching credentials are typically awarded, namely preservice credit hours, degrees, grade-point averages, in-service credit hours, etc. The omission is intentional and is based on the judgment that such factors are not valid indicators of teacher effectiveness. Their lack of validity stems from two sources: (1) nearly all teachers have been selected on such variables; hence in the selected group the restricted ranges of the variables work against their showing any correlation with effectiveness (e.g., intelligence and income are virtually uncorrelated among the alumni of the Harvard Graduate School of Business, though they are not uncorrelated in the general population); (2) many such "credentialing" factors have no direct effect on teaching effectiveness.

The argument that the bases of professional credentials are uncorrelated with teacher effectiveness because of the selectivity of the teaching profession doesn't jibe with the long-standing criticism of the profession's lack of selectivity. A "homogeneous clustering" phenomenon within school districts could account for both states of affairs. Suppose the qualifications of teachers are relatively similar within districts and relatively different across districts. This clustering pattern is highly plausible, considering current teacher selection practices and the economics of schooling. Furthermore, it is consistent with both the nonselectivity of the credentials-granting process and the absence within districts of a correlation between teacher effectiveness and the credentials criteria (cf. Figure 3).

I have no data on these questions of selectivity either within school districts or across the profession. Nonetheless, arguments such as those presented here lend credibility to the scattered empirical studies which show zero correlations between credentials criteria and measures of teacher effectiveness.

Rosenbloom et al. (1966, p. 108) concluded from a study of high-school mathematics teachers that "the effectiveness of teachers using the SMSG materials and measured by student learning is not influenced to any significant degree by the length of the teacher's experience in teaching mathematics, his undergraduate and graduate courses and grades, and his participation in professional mathematics organizations. Apparently if a teacher meets acceptable qualifications in

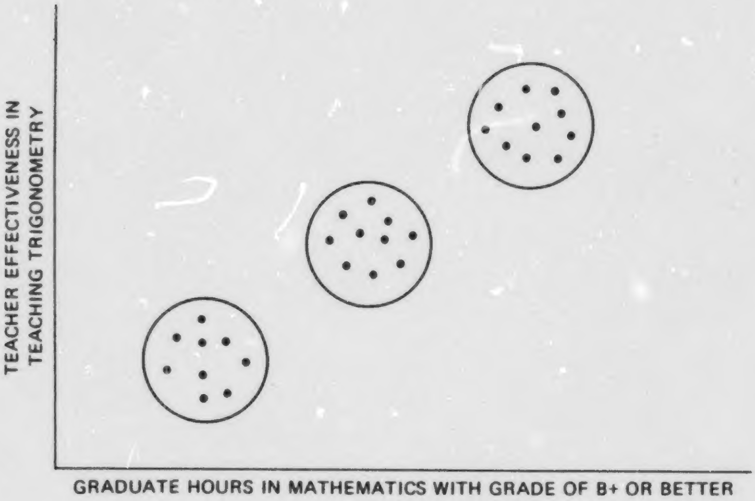

Figure 3. Hypothetical scatterplots of teacher training and pupil performance within and across districts.

these respects, higher qualifications do not make a difference."

One exception to the ban on "credentialing" data seems justified. It seems obvious (particularly at the high school level) that there are minimum levels of subject-matter knowledge below which teachers are incompetent to teach. (There are arguments for the instructional value of having an ignorant teacher, but even when such arguments can be made to stick, it's hard to justify paying ignorant teachers high wages.) Infrequently, a teacher's own grasp of the subject he is teaching drops below one of these minimum levels; we may, rarely, find the football coach covering Algebra III who can't graph a linear equation, or the choral music teacher who can't read music. Such instances are, no doubt, so rare that correlational studies of teacher knowledge and student learning continue to show zero relationships.

Nonetheless, they are present in each of our personal experiences and are sufficiently repugnant that they deserve attention in a teacher evaluation system, provided the costs are reasonable. Costs for this kind of evaluation can be held to a reasonable level by centralizing the knowledge testing of high-school teachers in the state education agency. The state department could produce short, "peaked" tests of knowledge in a dozen subject areas at low cost by using data from existing standardized tests for pupils. The tests could be developed, distributed, and scored by the state department for use only for local purposes.

Such an examination could be administered every other year or less often, instead of only once at entrance into the profession. Thus it would exercise some control over shifting teaching assignments that make for real administrative convenience but for no academic sense. Once again, the purpose of the testing would be to detect the one-in-a-thousand egregiously indefensible teaching assignment, not to grade or rank teachers on their subject-matter knowledge. The program would have a very modest goal; if its costs per teacher per year exceeded 10 cents, it probably wouldn't be worth maintaining.

Some of my students noted the absence of "peer-evaluation" recommendations in earlier drafts of this paper. I considered such techniques, but rejected them; the idea suggests a jury system in which the jury accepts the punishment if the defendant is found innocent and goes free if the defendant is found guilty. Nevertheless, I respect the system of peer evaluation to which I am subjected in my own university. My students got me to acknowledge this and my own condescension in considering my colleagues capable of fair and valid peer evaluation, but not high school teachers. I'm willing to rethink the notion of peer evaluation for the

Stull Act, but time has caught me short.

Postscript

Earlier drafts of this paper carried more colorful titles: "Act One," "Acting Up," "Acts, Romans, and I Corinthians," "Cleavers, Hatchets, and Other Acts," and "Stullen oder Hungertod?" This last, of course, is German, and it means "Bread-and-Butter or Starvation?" The choice it poses reflects my own ambivalence about accountability. The movement to make someone accountable for America's schools satisfies a broad range of motives: some people seek retribution; others blatantly act out their hostility against the Establishment--any establishment; still others find sources of renewed enthusiasm in accountability for their interest in children and their development. I suspect all of these motives in myself and in many with whom I talk. These machinations we're all going through in the name of accountability have many motives and more meanings than we realize to the people caught up in the swirl of activity. Systems analysts are designing accountability systems, but they're also somewhat grandly displaying their expertise. School superintendents are taking initiatives, but are the actions meant to say to a higher body, "Look, I'm cooperative and efficient and can be counted on in a more responsible position," or are they intended to say to the local community, "You can count on me to protect local autonomy?" Teachers are obediently recording each child's most insignificant acts, but privately they'll confess that they think it's nonsense.

It is shortsighted to view accountability and the forces it has released as having to do solely with fiscal responsibility or improved pupil learning. For some inexplicable reasons, accountability has surfaced as a celebrated cause, as only a few causes do in any decade. Accountability doesn't deserve the attention it is receiving--not if it consists of the feeble plans and devices for allegedly improving schools that I've heard presented in the name of accountability. Accountability owes its current prominence to more powerful social forces than those of empirical educational research and evaluation.

I have vague, private feelings that accountability reflects a fundamental struggle between those who possess some degree of trust in the developmental regularities of social and human organisms and those who trust only their own power to manage other people's comings and goings.

I believe that there exist a few people who know what is best most of the time for most other people. But when these few people achieve the power to act on their wisdom, their inevitable mistakes can be catastrophic, because of the scope of that power. The mass of men of less wisdom acting alone make mistakes more often, but fewer people are hurt and the damage is less serious. And this has something to do with the Federalist Papers, and American history, and the schools of the 1970's.

References

Anderson, R. C. How to construct achievement tests to assess comprehension. Review of Educational Research, 1972, 42, 145-170.

Bane, M. J., and Jencks, C. The schools and equal opportunity. Saturday Review, 1972, 55, No. 38, 37-42.

Belgard, M., Rosenshine, B., and Gage, N. L. Explorations of the teacher's effectiveness in explaining. In Westbury, I. and Bellack, A. A. (Eds.). Research into classroom processes: Recent developments and next steps. New York: Teachers College Press, 1971.

Bennet, R. W. The magnitude of the teacher effect upon pupil achievement in the primary grades. Doctoral dissertation, University of Colorado, Boulder, Colorado, 1971.

Bormuth, J. *On the theory of achievement test items.* Chicago: University of Chicago Press, 1970.

Brophy, J. E. Stability in teacher effectiveness. *R & D Report Series, No. 77*. Austin, Texas: Research and Development Center for Teacher Education, University of Texas, 1972.

Colver, R. M. Estimating item indices by monographs. *Psychometrika*, 1959, *24*, 179-185.

Connor, A. Cross-validating two performance tests of instructional proficiency. University of California, Los Angeles. U.S.O.E. Project No. 8-1-174, 1969.

Cronbach, L. J., and Furby, L. How should we measure change --or should we? *Psychological Bulletin*, 1970, *74*, 68-80.

Forsyth, R. A., and Feldt, L. S. An investigation of empirical sampling distributions of correlation coefficients corrected for attenuation. *Educational and Psychological Measurement*, 1969, *29*, 61-71.

Fortune, J. C. A study of the generality of presenting behaviors in teaching preschool children. Memphis, Tenn.: Memphis State University, 1966.

Fortune, J. C. A study of the generality of presenting behaviors in teaching. U. S. Office of Education Project No. 6-8464, Memphis State University, 1967. (ED 016 285)

Fortune, J. C., Gage, N. L., and Shutes, R. E. The generality of the ability to explain. Paper presented to the American Educational Research Association. Amherst, Ma.: Author, University of Massachusetts, February 1966.

Gage, N. L. (Ed.). *Handbook of research on teaching.* Chicago: Rand McNally & Co., 1963.

Glass, G. V, Booth, D. J., Collins, J. R., Erion, J. G., Horn, J. G., James, H. H., Peckham, P. D., Remer, R., Shepard, L. A., and Wing, D. R. W. Data analysis of the 1968-1969 Survey of Compensatory Education

(Title I). Final Report, Grant OEG-8-8-961860, 4003-(058), U. S. Office of Education, September 1970.

Harris, A. J., et al. A continuation of the CRAFT project: Comparing reading approaches with disadvantaged urban Negro children in primary grades. Cooperative Research Project No. 5-0570-2-12-1. Division of Teacher Education of the City University of New York, 1968. (ED 010 297)

Justiz, T. B. A method of identifying the effective teacher. Doctoral thesis, University of California, Los Angeles. Ann Arbor: University Microfilms, 1969. No. 29-3022-A.

Kaplan, A. The conduct of inquiry. San Francisco: Chandler Publishing Co., 1964.

McClellan, J. E. Classroom-teaching research: A philosophical critique. In Westbury, I., and Bellack, A. A. (Eds.). Research into classroom processes: Recent developments and next steps. New York: Teachers College Press, 1971.

McNeil, J. D. Performance tests: Assessing teachers of reading. The Reading Teacher, 1972, 622-627.

Medley, D. M., and Mitzel, H. E. Measuring classroom behavior by systematic observation. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally & Co., 1963.

Millman, J. Performance-based indicators of teaching effectiveness: Has the time come? Paper presented at the annual meeting of the Northeastern Educational Research Association, November 1971. Ithaca, N. Y.: Author, Cornell University.

Popham, W. J. Designing teacher evaluation systems: A series of suggestions for establishing teacher assessment procedures as required by the Stull Bill (AB 293), 1971 California Legislature. Los Angeles: The Instructional Objectives Exchange (Box 24095). 1971(a).

Popham, W. J. Performance tests of teaching proficiency: Rationale, development, and validation. American Educational Research Journal, 1971(b), 8, 105-117.

Rogers, W. T. Jackknifing disattenuated correlations. Unpublished doctoral dissertation, University of Colorado, 1971.

Rosenbloom, P. C., et al. Characteristics of mathematics teachers that affect students' learning. Cooperative Research Report. University of Minnesota, 1966. (ED 021 707)

Rosenshine, B. The stability of teacher effects upon student achievement. Review of Educational Research, 1970, 40, 647-662.

Rosenshine, B. Teaching behaviors related to pupil achievement: A review of research. In Westbury, I. and Bellack, A. A. (Eds.). Research into classroom processes: Recent developments and next steps. New York: Teachers College Press, 1971.

Ross, J., and Weitzman, R. A. The twenty-seven per cent rule. Annals of Mathematical Statistics, 1964, 35, 214-221.

Scriven, M. A summative evaluation of the products of the Experimental Schools evaluation meeting. In Glass, G. V, Byers, M. L., and Worthen, B. R. Recommendations for the evaluation of Experimental Schools projects. Final report, USOE Grant #OEG-0-9-152075-3720. Laboratory of Educational Research, University of Colorado, 1972.

Stake, R. E. Testing hazards in performance contracting. Phi Delta Kappan, 1971, 52, 583-588.

Walberg, H. J., and Anderson, G. J. Classroom climate and individual learning. Journal of Educational Psychology, 1968, 59, 414-419.

Walberg, H. J. Predicting class learning. An approach to the class as a social system. American Educational Research Journal, 1969, 529-542.

DISCUSSION OF PAPER BY GENE V GLASS

HENRY M. LEVIN
Stanford University

We are very fortunate to have received this paper so early in the conference. I think it is an excellent paper and raises the right issues, even though perhaps it doesn't directly attempt to do that. It seems to me that over the next two days there is going to be a struggle between two elements of analysis regarding the Stull Act. One is just how we comply with the Act, and the other one is the larger contextual set of philosophical questions that face us when we deal with accountability. It is not clear at all that the Stull Act has very much to do with accountability, even though it has been interpreted in that manner, and I guess that that is what troubles me about Dr. Glass's presentation.

First let me state that I see no deficiencies in any of his statistical analysis and I agree with much of his philosophical analysis. But I'm still troubled because I see that most of you are, first of all, people who have to make decisions that comply with the Stull Act legislation. That suggests to me that the message of Dr. Glass's paper is really this--all of the mechanisms for trying to tie attributes of teachers to measures of student gain have enormous deficiencies. These mechanisms show very little stability from year to year over subject matter and so on. I think we should acknowledge this, and in fact I would be very interested to find any serious contradiction of what Dr. Glass said on this score. But on the other hand, the schools have to respond to some legislation in the here and now. If we are told that somehow we have to go through the motions of curing cancer even though there is no cure, the necessities of life suggest that we have to give placebos or something similar just to fulfill those requirements. And this, it seems to me, is the message here: "Go through the motions, do what you have to do to satisfy the stringencies of the Act, but at the same time be aware both of what you are doing and of what you are not doing." I think in this context the criticism of the various approaches is legitimate and important. That is, to fulfill the motions of the Stull Act is one matter and (one would hope) to want to fulfill the spirit is another; but the present state of the art not only of statistical methodology but also of accountability methodology is too little developed for us to believe that we are actually producing an accountability mechanism. This, it seems to me, is the message that should come out of Dr. Glass's paper on statistical analysis.

Now let me mention two other matters that I think are important. First, in the actual implementation of the Act, there is again a schism between the practical and the philosophical approaches. On the practical side, one point that troubles me is that there is a difference between a statistically significant finding, reliability coefficients, stability coefficients, and the kind of apparatus we need to secure justice. If, for example, we have a mechanism we can all agree upon that will separate those teachers who are "acceptable" from those who are "not acceptable" and if we can be certain at the 95% confidence level (that is, 19 out of 20 times), the choice would not be made at random. We are still saying, and this corresponds to Dr. Glass's analogy, that we will be making a mistake in judgment about one of those teachers--we will be putting him into one group when he should have been put in the other group. We have about 200,000 teachers in California, so the suggestion is that, in the simplest sense, if we were to rate all of them, about 5% (around 10,000 teachers) would have an injustice done to them, even though according to the probability law of large numbers our statistics would have been correct, that is, assuming that the state of the art is a lot better than it is now and that we can agree on what should be measured, how it should be measured, and so on.

We must consider that the kind of statistical technology and the approaches we use in evaluation may not be very appropriate for this kind of accountability question, where the result may lead to a teacher's being released, when in fact it is just a statistical artifact in perhaps one out of twenty cases. Let me state my own prejudice on this issue. I don't believe that accountability is a statistical problem. I don't even believe it is a technical problem. I think that we as educators make it a technical problem because this is how we get our bread and butter. I get paid consulting fees not for saying accountability is a philosophical or a political problem but for saying it is a technical problem, because people who use my services aren't willing to pay for my political views or my philosophical views, but they are willing to pay for my technical views. Therefore, many of us are fond of seeing all educational problems as essentially technical problems, and we use highly technical terminology. But I would assert that what we are really talking about is a political problem--one that deals with the value perspectives of a society. If lower-class children, wherever they are, are being socialized to become lower-class adults (and the evidence with which I am familiar suggests that is true), this also suggests that the entire society is doing that; that is, this is part of the neighborhood, part of family socialization practices, and is related to the kinds of schools children attend. Then, I'd question whether we should even be worrying about whether individual teachers are doing relatively well or not well with individual students, particularly if the kinds of outputs we are talking about do not appear to be very important with regard to the eventual social reality and the eventual economic status of students.

Many of you may have been shocked or disturbed by the article by Bane and Jencks that was mentioned by Dr. Glass, and I am sure you are going to be even more shocked when you read the book on which that article is based. I think there are a lot of grounds for criticism of that book, but it seems to me Jencks does ask the right question, and that is, what aspects of schools, or homes, are associated with success in adult life?

One of the surprising results is that although there is a statistically significant relationship between cognitive outcomes and such measures of success as occupational status and economic status, the correlations are terribly small. In contrast, recent literature has suggested that noncognitive aspects are probably more important than cognitive ones, something which was not discussed directly in the statistical evaluation of accountability. And indeed if you look at lower-class schools and so-called middle-class schools you see very different patterns of social life. In some upper-middle-class schools, for example, the students are free to walk off or on campus whenever they want. They decide their own curricula and argue with the teachers with impunity. They make many of the decisions that affect their lives in that context. They are learning to choose among alternatives. On the other hand, in most lower-class schools students are learning something very different. They are learning, first of all, that even to satisfy nature's urge they must raise their hands, and even then the teacher can be punitive and not give them permission to leave the room. If they are caught in the hall without a note from the teacher they are in trouble. They do not have curriculum choices--the choices are made by paternalistic counselors and others who "have their best interests in mind." They are subject to a high degree of discipline. And out of this system we see low cognitive attainments, but we also see a very distinct mode of noncognitive behavior. We see that these are students who, having never had choices among alternatives, are not able to make such choices in their consumer behavior, in processing information about jobs, and so on. These students make ideal blue-collar workers because they can take orders very well.

On the other hand, the children from the upper-middle-class schools seem to make very good bosses and managers. They seem to be able to give orders, they seem to be able to choose among alternatives, they seem to be able to evaluate complex information and come up with answers. Now I assert that this is a problem of the society, and that when we talk about such cognitive measures as which teacher is doing "better" than which other teacher, or which teacher is doing poorly, we are ignoring the big thing. The big thing is that on the average the best predictor of cognitive or noncognitive outcomes in life is social class. And even if one teacher can do somewhat better than another teacher with lower-class students, the point is that the schools or all of our social institutions don't do a good enough job to make the accident of birth meaningless in the race for life's rewards. I want to leave you with that point of view, to remind you how easy it is to get bogged down in talking about the statistical aspects of accountability and simply become the hand-maidens of the political forces who have created a system in which the large questions of accountability are ignored.

HENRY F. KAISER
University of California, Berkeley

Professor Glass discusses three methods of evaluation in his paper: (1) standardized tests; (2) the so-called PM^2 method; and (3) the observational-judgmental system (his own suggestion, and perhaps the most appropriate method). He rather quickly got rid of standardized tests as being "obviously" inappropriate. I'll spend even less time on them here, but I would like to come back to them.

It seems to me that, with respect to the second method --the PM^2 method--Professor Glass has presented rather overwhelming evidence that, at least as presently developed, it is not sufficient to the task. In the PM^2 method one really has a very small sample of teacher behavior, and, as Professor Glass pointed out, this sample may not be a very good one. One would hope that a teacher has a good knowledge of his subject matter, and this may not be the case with some of the structured tests such as those suggested in the PM^2 method. Going along with a note that Dean Coladarci made* about my being too concerned with reliability, I'll plead guilty. The reliability figures for the PM^2 method, as presently developed, are really depressingly small, and reliability is a necessary, if not sufficient, condition for good measurement. Thus reliability certainly would have to be improved. Of course, one can always improve reliability if he takes the time to do it, but unfortunately it doesn't go up linearly with time expended. Thus, to make the PM^2 method a viable approach, it appears that a tremendous amount of data would have to be collected. The thought crossed my mind--I'm not completely a theoretician--that the PM^2 method would be very expensive to administer.

---

*In comments from the floor, not included in these proceedings.

Let me turn next to the observational-judgmental system proposed by Professor Glass. To a lesser extent some of the same problems do arise. This system calls for a very small sample of teacher behavior and maybe not a very representative one. The teacher would probably be "uptight" when being observed, as opposed to being a smoothly functioning teacher in other instances. However, the reliability here seems a little better than in the PM^2 method. This is just the impression I have, but perhaps I am being biased by Professor Glass's presentation. One negative thought I had about Professor Glass's suggestion is the same thing I just mentioned with respect to the PM^2 method (and it seems to me that this might hold even more strongly here). That is, the system would be very expensive to administer because it would require a large corps of competent observers. We have some 200,000 teachers in California, and to observe all of them would be a costly enterprise.

I wish Professor Glass had explained at greater length why the use of standardized tests is inappropriate, because, in all my innocence, that is where I would start. I am concerned with reliability, and although these standardized tests might not measure exactly what we want to measure, what they do measure they measure well. Thus they might be worthwhile. With standardized tests which are routinely administered to children, it would seem, practically speaking, that the teacher would not be disrupted and there would not be the problem of a small sample of teacher behavior.

I haven't read Bob Stake's article,[*] but where Stake points out that standardized tests are not valid for this purpose, I can't help but wonder if they might not be made valid if sufficiently hard effort were put forth by sufficiently clever people. The costs of using standardized tests, since they are routinely administered for other purposes, wouldn't be particularly great. I suppose one would have to hire some very clever people to handle the statistical data. To repeat, what I am saying is that I am attracted by the extremely high reliability of individual achievement test scores.

A final thought is that perhaps Professor Glass's suggestions could be combined with the use of standardized tests in some fashion. With 200,000 teachers in California, we perhaps could take a sample of, say, 1,000 and have them evaluated observationally-judgmentally with great care by highly trained people. Then we could predict this judgment with standardized achievement tests through standard procedures, multiple regression or, if the judgmental scores were multivariate, we would have a canonical correlation problem. And perhaps get, or be able to predict, the criteria very well. If in fact we were to do this, that is, predict classes of observational-judgmental scores using a small sample from the total population, we could then use the prediction equations to predict the observational-judgmental scores for all 200,000 teachers.

I can see problems in doing this, of course. One thought that comes to mind is that we might, or might not, want to partial general intelligence out of the system. This is something that would have to be determined empirically, at least initially. In fact, if the prediction were efficient, we would then be faced with the problem of trying to make sense out of the variables we had developed. That has always proved to be a tough problem, but it seems that it might yield eventually, again to sufficiently clever people.

---

[*] R. E. Stake, Testing hazards in performance contracting, Phi Delta Kappan, 52(1971), 583-588.

W. JAMES POPHAM

University of California, Los Angeles

I arrived a little bit late for Dr. Glass's presentation, so for a moment I would like to address myself to some of the remarks made in the last part of his morning session. I am pleased that some of my work has drawn the attention of Dr. Glass. I respect him as one of the best research methodologists in the nation, and therefore to have him direct his criticism toward our work is gratifying. I suspect that we shall profit from his analysis.

Dr. Glass is distressed generally with product measures and with a teaching performance test in particular, and I'd like to deal with some of his criticisms. Frankly, I think they would be appropriate if one were proposing that teaching performance tests at this moment represent a refined measurement technique that ought to be employed on a wide-scale basis in situations such as that faced by California educators trying to implement the Stull Act. But this is not the contention of most of the people working with these measures. It is extremely early in the game for a teaching performance test. The first paper read at a research association describing such measures was presented less than six years ago--and in the history of any kind of technical development of a measurement, that's awfully early. And if I have, in my enthusiasm for trying to counter process-oriented evaluators, attempted to suggest that teaching performance tests could be a useful tool, I certainly did not mean to suggest that the technology currently makes it possible to use these measures with confidence on a wide-scale basis.

Dr. Kaiser suggested that with hard work by bright people we might be able to improve the quality of standardized tests. I am very encouraged about the prospect of improving the quality of teaching performance tests if we secure hard work by good people. The fact that Dr. Millman at Cornell University has recently participated in our efforts to test precisely the questions that concern Dr. Glass--namely, the reliability of these measures--is, I think, encouraging. We had hoped very much to have some data analyses to report, based on studies run last year. We were investigating not only the relationship among somewhat dissimilar types of teaching performance tests but also the relationship among scores of people taking the same performance test with different learners. I think that that was Dr. Berliner's point this morning; when an individual is engaged in two discernibly different performance tests, then perhaps it is not appropriate to view them as bearing on test-retest reliability. Rather, it is appropriate to view them as different kinds of measures, and one would expect, therefore, weaker correlations among them. But I do not pretend that this particular approach has received adequate scrutiny or strong support; rather, I think it is potentially a good idea, as Millman says.

When Dr. Glass makes alarming predictions regarding the cost of using this kind of approach, based on reliability estimates on these very early and terribly primitive studies, I think it is worse than a lack of generosity--I think it is a lack of fairness. I believe that this is a potentially useful avenue for inquiry; at this moment it is far too early either to applaud it or to write it off.

I suspect that the teaching performance tests (and Dr. Glass concedes this point) may at the moment have some utility in connection with promoting improvement in instruction. Most of you involved in Stull Act implementation realize that a corollary segment of that legislation requires us to devise improvement schemes for teachers to help them get better at what they're doing. I believe that teaching performance tests, or instructional mini-lessons, focused on the products rather than on the process employed by teachers, may have some utility here. Dr. Glass, of course, pointed that out. Such tests may also have some value when it comes to trying to test the efficacy of an inservice program or a preservice program, to see whether we can improve teachers' skills in the kinds of situations where the focus is not on making a judgment about the individual teacher. I would hate to have my own children's teachers judged on the basis of very primitive teaching performance tests, but I would not be reluctant to judge my own success as a teacher educator on whether a group of my students is able to get better at that particular enterprise. One thing that I do with my classes at the beginning of each quarter is to test myself against the students on a brand-new performance test about which I have no previous information. I do this to show my students how I want them to practice on these performance tests, because I think it's useful as a model. But more important, I want to show them that I can beat them on the test, because I want to establish expert power. I want to show them that a person who is cognizant of rules regarding instructional techniques can outperform a novice. The effect thus far has been very dramatic. My students attend to what I'm saying--it's no longer the teacher education pap that is so typical of preservice courses. Rather, they see that there is in fact something one can employ that increases one's ability to accomplish prespecified objectives.

This is only one critical technique by which we should judge teachers, but it is one criterion that can be potentially useful, I think, if we sharpen our technology. The point of view that I saw expressed differently in Dr. Glass's approach to this measurement dilemma and in my own deals with whether one attends primarily to products or to process. It seems to me one can divide most of the early Stull Act implementation schemes into those in which the attention is focused on what people do during instruction (and one looks at it as it's happening or perhaps asks students or peers or others to judge how it happened) and those in which one looks at the results of what the teachers do--the product of those instructional acts. I think there is a division with respect to what we should be primarily (but not exclusively) attending to--and I would primarily, but not exclusively, attend to results of instruction rather than to the process of instruction.

PROBLEMS OF DEFINING EDUCATIONAL OBJECTIVES AND STANDARDS IN IMPLEMENTING THE STULL ACT

W. JAMES POPHAM
University of California, Los Angeles

The architects of the recently enacted California teacher evaluation law left the implementers of that law with far more slack than is typical in the building industry, where architectural specifications must be both comprehensive and detailed. Whether, in the case of the Stull Act, this latitude represented political sensitivity to local educational prerogatives or merely reflected shoddy architectural planning remains to be seen. Motives aside, California educators now have a vast amount of required explicating to do if they are to implement the new law.

Since it's still largely a local option enterprise, with relatively few procedural constraints spelled out in the legislation, efforts should be made to sensitize California educators to the range of defensible implementation alternatives available to them. Awareness of different varieties of concrete can help the contractor lay a more lasting foundation.

This paper will focus on implementation problems associated with defining educational objectives and standards for the required district personnel evaluation systems. The analysis will be restricted to considerations of teacher evaluation, and will not include the evaluation of nonteaching certified personnel.*

The impetus for this analysis stems from two binding sections of the Stull Act which, rather than allowing local district implementation schemes, require local educators to include the following elements in their teacher evaluation procedures:

---

*Problems related to the evaluation of nonteaching certificated personnel are treated elsewhere in these proceedings, by Max G. Abbott.

1. Establishment of standards of expected student progress in each area of study and of techniques for the assessment of that progress.
2. Assessment of certificated personnel competence as it relates to the established standards.

There are two key terms, standards and progress, in these stipulations which must be inspected further. Dealing with the easier analysis first, the term standard shouldn't pose much difficulty in interpretation since the technical denotations and connotations of the term pretty much parallel popular usage. When most people use the term standard they refer to a level or degree of proficiency. In educational parlance the term generally describes how well we expect a person or a program to perform. Later in this paper an attempt will be made to consider alternative techniques for establishing educational standards.

The second term, progress, is more vexing. For some people it means pretest to posttest growth. For others it simply means how are the students doing. Some educators equate progress with the degree of objectives attainment. To some it merely represents the substance of whatever we apply standards to. The possibilities for disagreement over basic terms and for resultant confusion are rife with respect to the term progress. An operational definition is clearly required.

Although attempting to infer legislative intent from actual legislation is as risky as reading pharmacy prescriptions upside down from a mirror, it appears that the California legislators responsible for the Stull Act were using the term progress to describe what students were learning.

They reserved the term standards to get at how well the students were learning whatever it was they were supposed to learn.

Thus, rather than assuming that there is an automatic equivalence between "progress" and "instructional objectives," it makes more sense to examine alternative indicators of progress and then to see how instructional objectives might relate to such indicators.

Indicators of "Expected Student Progress"

By and large the indicators of student progress which will be employed to implement the Stull Act will be learner performance on measuring instruments. Although the measurement schemes will theoretically consist of a wide variety of assessment procedures, it is very likely that the measurement schemes initially selected will generally be simply paper-and-pencil tests. This will occur less because of the lack of imagination of Stull Act implementers than because of the astonishing imposition by the legislators of a statewide teacher evaluation enterprise without providing adequate financial resources to give that enterprise a decent chance to succeed. Financially frantic districts will be hard put to provide any other measurement mechanics than those which can be acquired or developed inexpensively.

Before considering several of the measurement acquisition options open to Stull Act implementers, it may be useful to remind ourselves that a local district can pursue at least four sources of measurement devices. First, there are commercial firms which distribute tests of various types. Second, if a district has sufficient financial resources plus measurement expertise, local development of tests is an eligible contender. Third, for affluent districts all or part of the test development effort can be contracted to one or more of the private consulting groups springing up to satisfy this type of need. Finally, a cooperative, interdistrict, test-development enterprise might be undertaken, for example involving a number of small or medium-sized districts within a particular county.

While there are other sources of good measuring instruments (truly desperate Stull Act implementers may wish to rub old lamps vigorously and then look for a large, turban-wearing stranger), those listed above represent the most common ways of acquiring the needed tests. In the next section of this paper, as we examine three different types of measurement procedures, remember that any of these acquisition routes can be employed: (1) purchases from commercial firms, (2) local development, (3) contracted development, or (4) interdistrict cooperative development.

Non-Referenced Measures

The treatment of alternative types of measurement approaches will be given short shrift in this paper because this topic is treated elsewhere* during this conference. Nevertheless, a word or two about norm-referenced tests is in order.

Generally speaking, norm-referenced tests, such as commercially distributed achievement tests, suffer from two major deficiencies which render them unsuitable for Stull Act implementation. First, they are by design so general (in order to serve an entire nation) that there is often little congruence between the emphases of the test and the emphases of a particular teacher (or school district). A mismatched measurement and instruction situation doesn't give the instructor a fair break.

---

*In the paper by Gene V Glass.

Second, certain psychometric requisites of norm-referenced tests, such as the necessity to produce considerable score variance, may lead to the elimination or modification of the very test items that would be most sensitive to detecting the instructional impact of a truly effective teacher. Thus, typical standardized achievement tests will penalize the high-quality teacher while perhaps not identifying the teacher most in need of improvement.

Finally, and I believe this is perhaps the most important deficiency of standardized tests for use in evaluating an instructional program (or an instructor), norm-referenced measures offer few cues to those teachers who, upon failing to perform satisfactorily, wish to improve themselves. For the vast majority of norm-referenced measures, there is no attempt made to delimit with precision the types of learner behaviors measured by the tests. As a consequence, the teacher only has a fuzzy notion of what the test really measured, other than that the students couldn't do it very well. An implicit assumption of evaluation programs is that we want to improve the efficacy of instructional efforts that are identified as inadequate. The ill-defined boundaries of most norm-referenced tests are not compatible with this assumption.

Criterion-Referenced Measures

In contrast to norm-referenced tests, advocacy of criterion-referenced tests is becoming very fashionable these days. One of the problems with this position, however, is that different people mean different things by the phrase "criterion-referenced measures." Robert Glaser, who has been instrumental in popularizing this distinction, believes that a criterion-referenced test measures the degree to which a student has achieved a given competency (often described in terms of an instructional objective). To others, a criterion-referenced test indicates whether a student has reached a given criterion level, that is, a standard of performance. These people very frequently hark back to classical prediction approaches in which the task was to forecast whether an individual would reach a given level of competence. To others, the criterion involved in a criterion-referenced test is not test behavior but some external, real-world behavior such as the individual's ability to cope satisfactorily with the price computation problems presented in a typical supermarket. And for still others, a criterion-referenced test simply refers to any kind of measure that is not a norm-referenced test.

These distinctions become important, of course, in that they influence how the evaluator proceeds with respect to the development and refinement of such measures. For purposes of this discussion, the interpretation supplied by Glaser* will be followed--a criterion-referenced test, we will assume, is designed to measure whether a learner has mastered a particular type of intended behavior.

The advantages of criterion-referenced tests for purposes of evaluation are several. In the first place, since they are not hobbled with the requirement to produce considerable score variance, but merely need to discern whether a given behavior falls within an individual's repertoire, the tests can more readily pick up the postinstruction behaviors that are achieved by most learners. Second, a criterion-referenced test is generally wedded to a more circumscribed description of learner behavior than is a norm-referenced test and consequently provides those evaluated with a better notion of what is actually being assessed. This increased clarity of description can lead to important instructional advantages. A key advantage of criterion-referenced tests is that typically they can be constructed so that they are more congruent with local curricular emphases. Because few such tests exist now, they will generally have to be devised locally and can therefore be tailored more directly to the instructional emphases within a given district.

---

*Robert Glaser, Instructional technology and the measurement of learning outcomes--some questions. In W. J. Popham (Ed.), Criterion-referenced measurement: An introduction, (Englewood Cliffs, N. J.: Educational Technology Publications, 1971), pp. 5-14.

Thus, the use of criterion-referenced tests seems to offer the teacher a better chance in Stull Act implementation schemes. These tests will generally be more sensitive to change and more consonant with local curricular foci. But there are some problems with criterion-referenced tests which should be recognized. First, if they are to be constructed locally (and this will frequently be the case), they will probably be generated by individuals who have less measurement sophistication than people who prepare commercially distributed tests. Furthermore, if the tests are constructed too narrowly, that is, so that only performance on a few test items rather than on a class of learner behaviors is measured, there is the danger that the teacher will instruct precisely to a test rather than for a range of eligible learner behaviors. Finally, because these tests will characteristically be "home grown," they may not induce the same degree of confidence on the part of the public that can be enlisted by standardized tests. In spite of these dangers, however, it still appears that for Stull Act implementation criterion-referenced tests would, in general, be preferable to their norm-referenced counterparts.

Domain-Referenced Tests

Although technically only a variation of criterion-referenced tests, the domain referenced test scheme devised by Wells Hively and his associates* has much to commend it for purposes of teacher evaluation. In Hively's approach, much attention is given to identifying the domain of eligible test behaviors assessed by a given measuring instrument. In other words, the measurement designer goes to great pains to clarify the eligible stimulus elements within that domain, the possible learner response options, the distinguishing criteria of correct and incorrect responses, the content limits to be included in the domain, etc. By thus explicating the nature of the domain to be assessed via a pool of test items, both the test developer and the individual being evaluated have a much more definitive idea of the boundaries of the measures involved.

Relationship of Alternative Indicators to Statements of Instructional Objectives

Just what is the relationship between the kinds of measures described above and statements of instructional objectives? As many California educators began this year to make implementation plans for the Stull Act, there was an obvious gravitation toward instructional objectives as a way of describing how students would display "progress." While this is only natural, considering the national emphasis on measurable instructional objectives during the past ten years, it is by no means necessary. Statements of objectives can, in a shorthand fashion, capture much of the meaning of a set of test items. In other words, rather than looking over a set of ten or twenty items designed to measure a particular kind of mathematics skill, the inspection of a well-formed instructional objective can save the educator much time in determining what is really involved in the test. Since educators are beginning to use objectives more frequently for curriculum decisions and instructional planning, it is only natural that statements of objectives will often be used to conceptualize the nature of "expected student progress."

---

*W. J. Popham, Teacher evaluation and domain-referenced measurement, in Educational Technology (in press).

But consider for a moment the three main operations of education, curricular decision-making, instructional decision-making, and evaluation decision-making. Objectives can prove useful in the first two of these without measurement devices, but when it comes to the evaluation operation, use of objectives without measurement is meaningless. Accordingly, in evaluation enterprises such as those designed to implement the Stull Act, measurement procedures must be constructed. Objectives may or may not be used to describe those measures. This is the primary role of objectives in connection with the Stull Act--that is, they can serve as economical descriptors.

Problems of Defining Objectives

Assuming that statements of objectives may be used as a way of helping teachers and others involved in Stull Act implementation define the nature of student progress, there are clearly alternative ways in which a useful set of objectives can be derived. Some of these should be examined. It must be noted that while the language of the Stull Act emphasizes the assessment of learner progress in "each area of study," implementers of the new law should not over-look those easily ignored but terribly important affective goals that most of us hope our schools are promoting.

Local Generation of Objectives

One procedure for the identification of objectives which has been followed by many school districts in the past decade is to have district teachers and administrators participate in the development of measurable objectives suited to that district. The chief benefits of this approach are that (1) the objectives can be explicitly tied to the curricular emphases of the district and (2) local educators feel they have a vital role in determining the objectives.

The disadvantages, however, appear to outweigh the advantages. In the first place, many teachers and other district education personnel are not particularly expert in the generation of objectives. Hence, one can anticipate a fair number of ambiguous, nonmeasurable goals as well as more than the necessary number of low-level, trivial objectives. Furthermore, most district educational personnel are already far too busy with their normal educational responsibilities to spend much time in the generation of objectives. Finally, there is something intrinsically debilitating in realizing that you are engaged in a wheel reinvention enterprise, for since so many measurable objectives have already been generated in various parts of the country for such evaluation efforts, it seems unfortunate to ask any individual to create anew that which has already been born.

Interdistrict Sharing

A way of partially avoiding the problems associated with local generation of objectives is to engage several districts, perhaps a whole county, in the task of generating appropriate objectives. Even assuming that there would be some differences among districts with respect to instructional emphases, these could be more readily accommodated thereafter by a local district. The advantage of pooling resources in this manner, of course, is that it reduces inadvertent duplication of effort and can capitalize on the most capable representatives of several districts.

The Use of Objectives Depositories

In the past few years several objectives depositories such as the Instructional Objectives Exchange[*] have been created which make available to educators collections of measurable objectives from which selections can be made for local purposes. It seems clear that, if objectives are to serve as the rubric around which Stull Act implementers will organize their thinking, an inspection of the objectives available from such depositories should be undertaken.

Technical Difficulties

It would be unfortunate not to alert Stull Act implementers to several technical problems associated with the use of objectives which have not been satisfactorily resolved. One of these deals with the level of specificity at which the objectives should be written. One can conceive of extremes along a continuum of specificity in which, at one end, we find a highly explicit objective that is essentially equivalent to a test item (the student will list the names of six U.S. presidents) and, at the other end, an extremely imprecise objective written in nonbehavioral form (the student will appreciate the importance of good citizenship). Unfortunately, the appropriate point between these two extremes has not been identified. We do not currently possess any techniques for replicably generating precise levels of specificity for measurable objectives. Those districts involved in the selection or generation of objectives must anticipate considerable latitude in the degree of specificity associated with those objectives, even though they be measureable.

---

[*] Instructional Objectives Exchange, Box 24095, Los Angeles, California, 90024.

A second problem deals with the appropriateness of the objective selected. Although we might wish for more definitive guidelines regarding whether the objectives chosen for such evaluation operations as the Stull Act are, in fact, appropriate, these kinds of decisions still rest largely on personal judgment. We don't yet have a defensible technology for deciding what ought to be taught, even though the use of objectives clarifies far better than any previous technique exactly what the nature of a goal intention really is.

These two problems only illustrate the point that whereas the use of measurable objectives adds much to the promotion of clarity, the technical difficulties associated with using such objectives statements are considerable. Stull Act implementers who employ objectives may be better off than those who employ nebulous goal statements, but the techniques of formulating objectives statements are not as advanced as we might wish.

Needs Assessment Strategies

Once it has been decided that objectives can prove a serviceable descriptor for Stull Act implementation, the question arises as to how decisions will be made with respect to what the final objectives really ought to be. There have been several attempts in recent years to tackle this problem in a technical manner through the use of educational needs assessment approaches. According to this general strategy, preferences are secured from a variety of individuals associated with the educational enterprise--for example, students, community representatives, and teachers. Either by ranking or rating extant sets of objectives or by generating new goals themselves, different inputs are secured from various representative groups. These preferences are then translated, sometimes by highly quantitative techniques, into overall indicators regarding the relative priorities of different objectives. The technical procedures for employing such multigroup preference determination schemes are outlined in some detail elsewhere.*

Summary

In review, if instructional objectives are to serve as descriptors for thinking about Stull Act measurement operations, a local district may secure such objectives by one or more of the schemes described above and may then decide in some way which of those objectives are most important.

Care should be taken when adopting objectives as a descriptive vehicle not to generate so many objectives statements that the result is an intellectually unmanageable collection. In 1971, for example, the Wisconsin legislature enacted a law requiring an annual statewide assessment. Designers of the plan to implement that law have adopted statements of instructional objectives as the descriptive language for conducting the assessment. Yet in the field of mathematics (grades K-8) they are now working with somewhat in excess of 400 objectives and will be assessing student performance at the end of the third grade with more than 300 of those objectives. Now, most human beings simply cannot comprehend the educational meaning of 300 disparate objectives. Although no empirical studies have been conducted on this point, one suspects that being able to cognitively process many more than 15 or 20 objectives per academic year is quite unlikely, and that that number probably represents sufficient detail for most educators. As a consequence, objectives must be stated at a greater level of generality, and perhaps only a few objectives should be emphasized.

---

*R. E. Stake, Priorities planning: Judging the importance of objectives, Instructional Objectives Exchange, 1972; W. J. Popham, An evaluation guidebook: A set of practical guidelines for the educational evaluator, Instructional Objectives Exchange, 1972.

Deciding on Defensible Standards

We now turn to the super-sticky problem of determining the standards of expected student progress against which teacher competence is to be appraised. Here, of course, the quantitatively oriented methodologist starts to perspire a bit, for there will have to be something other than numbers and test scores involved--there will have to be judgment. This is so because, no matter how many numerical machinations we employ during the process of determining how well we want youngsters to perform, there will necessarily be abundant doses of personal estimates regarding how much to weigh various factors in order to spell out a desired level of learner proficiency.

A fair amount of discussion has been carried on in evaluation circles recently regarding the merits of a distinction between relative and absolute standards. Perhaps such discussions have merit at a theoretical level, but from the practitioner's plateau it is hard to visualize a decision about where to establish an educational standard that is not heavily dependent on experience-based comparisons. For example, once a given type of learner behavior has been isolated, our notions of how well we want learners to display that behavior are contingent on previous experience regarding how well we or others should be able to do with respect to that behavior under typical or possibly ideal conditions. How perplexing it would be if someone defined an innovative type of cognitive skill for learners, one which bore little relationship to skills with which we were familiar, and then asked us to specify how well 12-year-old children should be expected to perform the skill. Unless we had some experience to call on in conjuring up performance standards, we'd probably draw a blank. For several of the procedural suggestions to be offered in the following paragraphs, a conscious effort has thus been made to bolster the decision maker's experience base.

Another general point needs to be made regarding standards, one so obvious it almost need not be mentioned. This is that realistic standards must take into consideration the entry behaviors of the learners involved. Clearly, for example, we would expect a different level of proficiency in reading for first graders than we would for high school seniors. Just as clearly, youngsters from an environment in which the merits of intellectual achievement have systematically been praised should be expected to perform at a different level of proficiency than we should expect in children from an environment in which intellectual achievement is ignored or denigrated. There should not be one standard--statewide or nationwide--for children at a given point in their lives. Defensible levels of expectation must involve the tailoring of standards to situational variables and personal antecedents.

This raises an interesting question regarding diversity of standards: should we lower our expectations for educationally disadvantaged youngsters to the point where we will require no more of them than second-class studentship? Definitely not! What we may have to do is move more slowly toward desired standards of excellence. Whereas a fourth grade teacher may not be evaluated adversely if his or her ghetto-background students are reaching only 25 percent on a given skill dimension when children of the same age from an economically advantaged suburb are reaching 50 percent, we cannot be satisfied with the proficiency levels of the disadvantaged children. To correct such deficits will undoubtedly require either far greater expenditures of the nation's resources for education or a major reallocation of current resources. In the meanwhile, in terms of evaluating a teacher according to specific standards of learner progress, the standards must be predicated on what the learners bring into the instructional situation.

Let's examine some alternative procedures for deciding on the kinds of standards that might be involved in a Stull Act teacher evaluation program.

The Estimated Percentage Method

One rather common procedure which dates back to the early days of programmed instruction involves having the evaluator make a judgment regarding what percentage of the learners should perform with what percentage of proficiency. An expectation of a 90-90 percentage proficiency standard might be set for a given test which signifies that at least 90 percent of the learners are expected to earn scores of 90 percent or better on the test. Often these percentage estimates are made almost arbitrarily, sometimes on the basis of more systematic data. But generally this procedure requires the estimator(s) to make pretty intuitive judgments regarding expected levels of learner proficiency.

A Needs Assessment Analog

As described earlier in the paper, one of the identifiably new activities in educational planning during recent years is the one referred to as "educational needs assessment." Stemming largely from a requirement in the distribution of ESEA Title III funds which stipulates that each state must conduct such a needs assessment, these operations have become well known throughout the nation. Many states have chosen needs assessment strategies which consciously attempt to be responsive to the preferences of groups with a legitimate interest in the educational enterprise. There are, of course, important procedural considerations, e.g., (1) should different groups' preferences be given different weightings in determining the final statement of needs, and (2) to what extent should data regarding learners' current status be used in the analysis. But regardless of the procedural variants, the key to these needs assessment strategies is that different groups are allowed inputs in the mechanism for the identification of educational needs.

It takes only a modest extension of the approach involved in goals or needs determination strategies to apply an analogous scheme to the determination of standards. Various groups could be apprised of the needs for such standards and could then be allowed to proffer their views regarding goals and regarding how well students should be expected to perform various tasks. While this approach might well involve estimated percentage techniques, it reflects a more systematic, diverse-input scheme.

A Simple Improvement Technique

Although apparently too simplistic for consideration, a straightforward scheme for deciding on standards might consist of merely expecting the progress of learners during a particular year to be better than it was during the preceding year. For instance, if a posttest-minus-pretest-equals-gain model is employed, a standard involving no more than the following could be used: gain this year greater than gain last year.

In the absence of more defensible standards-determination schemes, a simple improvement technique may prove helpful as an interim solution. It seems certain that whatever resources can be used by a district to implement the Stull Act will, at the outset, more likely be expended on decisions regarding what types of progress to identify and how to measure it. A simple improvement technique can serve as a satisfactory temporizing vehicle.

One of the practical difficulties associated with such methods is that, if there is no history of the previous performance of district pupils on certain measures, it is difficult to establish the baseline on which improvement can be predicated. One method of playing this kind of game is to use standardized test grade-level equivalents (i.e., one academic year's expected growth) as a guide, not a definitive quantity. By inspecting standardized tests themselves, as well as the accompanying normative data per grade level, some limited extrapolations can cautiously be made to guide a local district.

An Asymptote Option

One of the requirements of the Stull Act which has yet to receive the attention it warrants is a provision that each district's implementation plan incorporate an assistance program for the teachers being evaluated. In other words, the new law mandates not only a teacher appraisal system but also a staff development program. Assuming that a district is serious about this improvement requirement and does invest a substantial amount of resources in an effort to improve the effectiveness of its teachers, we might reasonably expect some kind of payoff, albeit modest perhaps, in staff effectiveness. This improvement would undoubtedly have to peak at some point, for it is unlikely that staff improvement would occur ad infinitum. Thus we might find, if a substantial staff improvement program is carried out each year, learner performance gains such as those seen in Figure 1. Now, at the asymptote of the improvement curve we could identify a level of expected learner performance based on more substantial evidence than mere guesswork.

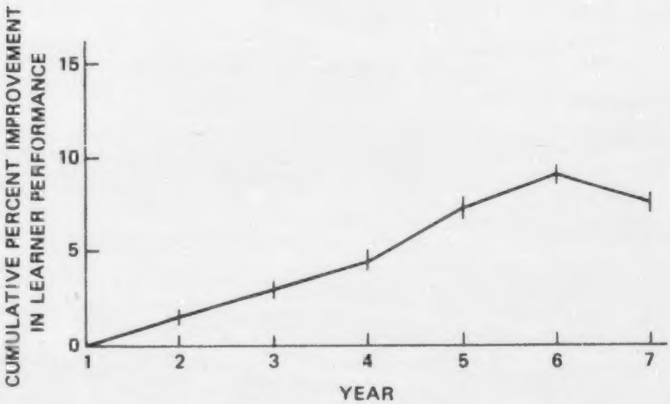

Figure 1. Hypothetical levels of improvement in learner performance based on a substantial staff development program.

Quite clearly, this approach represents little more than a refinement of the simple improvement technique. But in districts where a major effort to improve staff competence is underway, it is an option worth considering.

A Regression Solution

In a recent paper, Klein and Alkin have reminded us of a familiar but not widely employed procedure for establishing standards, namely, regression analysis.[*] This approach involves the administration of a pretest (any of the indicators of student progress described earlier) at the beginning of the year and a comparable or identical test at the close of the year. Based on the relationship between these two sets of data, an expected posttest score can be computed for each learner. By inspecting, teacher by teacher, the proportion of actual posttest scores which are equal to or above the expected score, we can calculate a per-teacher index of the degree to which the teacher is able to promote predicted (by the regression analysis) achievement. Because it offers a simple step-by-step explanation of this procedure along with easy-to-interpret examples, the Klein-Alkin article should be consulted by those interested in exploring this alternative further.[†]

Final Thoughts

In this paper I have attempted to restrict my remarks to the assigned questions dealing with standards and objectives. That proved to be a rather annoying assignment because, when dealing with a legislative enactment of this much potential import, the temptation is ever present to tangle with other procedural or philosophical issues. In other essays* I have attempted to treat some of these, but I have veered away in the present effort.

---

[*] S. P. Klein and M. C. Alkin, Evaluating teachers for outcome accountability, Evaluation Comment, Vol. 3, No. 3, May 1972.

[†] Copies of the Klein-Alkin paper are available from the Center for the Study of Evaluation, University of California, Los Angeles, 90024.

In many ways the assigned task has not been that entrancing. When I told a colleague some weeks ago that I had accepted this assignment, his response was: "What can you say that's new?" I didn't have an instant response to that question then. I have one now. The answer is--very little.

There are some mildly new wrinkles contained in the foregoing paragraphs, but nothing that would wrench the would-be Stull Act implementer from his seat with ecstatic cries of "The answer! The answer!" I guess that's because I believe the answer to objective and standards determination is rather prosaic. It's prosaic, and it involves hard intellectual work. The scenario for a typical Stull Act implementation drama would be something like the following:

Somebody has to decide in what areas of content and behavior learners will be measured to reflect teacher competence. Statements of instructional objectives are a handy way of talking about such learner performance. Then somebody has to decide on the kinds of measurement procedures needed to provide indicators of this performance. Next, somebody has to figure out a scheme for deciding how well students ought to perform, so that such standards can be contrasted with actual teacher performance. Finally, somebody has to compile all of these data and report them in an accurate and readily interpretable form.

I have attempted to set forth options, with some implicit and explicit preferences, for the chief phases of such an operation. Because the Stull Act, with all of its shortcomings, represents an effort to consider the worth of instructors in terms of learner results, I am supportive of the general intent of the new law. I hope that analyses such as those provided at this conference will help promote the kind of educational improvements anticipated by Assemblyman Stull and his colleagues.

---

*W. J. Popham, Designing teacher evaluation systems. Instructional Objectives Exchange, December 1971, Los Angeles, 90024; W. J. Popham, California's precedent-setting teacher evaluation law, Educational Researcher, Vol. 1, No. 7, June 1972.

DISCUSSION OF PAPER BY W. JAMES POPHAM

ELLIOT W. EISNER
Stanford University

Fifteen minutes is a brief period of time in which to respond to a paper dealing with an important piece of legislation, so I will proceed with less elaboration of the ideas I will present than I would have liked. Perhaps the most significant aspect of Dr. Popham's paper is that it accepts the premise of Assembly Bill 293 and is focused, therefore, on ways of implementing it. I do not share that premise and I would like in my remarks not only to respond to some of the points that Jim made in his presentation but also to raise some issues regarding the Act which I think are worth considering.

Let's start at the beginning by examining the context in which Article 5.5, dealing with the evaluation and assessment of performance of certificated employees, is found. This section of the Stull Act is, of course, only one section of an act that starts by describing the grounds upon which permanent employees of school districts--teachers, consultants, and administrators--can be fired from their jobs. That in itself is an auspicious way to begin a bill concerned with the evaluation of teaching effectiveness! The section describing the grounds for dismissal, it might be noted, is about five times as long as the section dealing with the mandate to develop uniform evaluation procedures in each school district in the state. The section dealing with grounds for dismissal states that no permanent employee can be dismissed except for one or more of the following reasons: immoral or unprofessional conduct, dishonesty, incompetency, conviction of a felony, and "knowing membership by the employee in the

Communist Party." Thus, at the very outset we find provisions for dismissal which are on their face a violation of constitutional rights. I deal with this section of the Act because I believe it sets the tone for that section with which this conference is concerned.

The Act is lacking in that the very purpose it is designed to serve is not stated. What is stated is: "It is the intent of the Legislature to establish a uniform system of evaluation and assessment of the performance of certificated personnel within each school district of the state. The system shall involve the development and adoption by each school district of objective evaluation and assessment guidelines." But to what end? Is the Act intended to improve the quality of teaching? Is it intended to provide evidence to the public that teachers are effective? Is it intended to be a tool to dismiss the incompetent? Is it intended to improve the quality of learning in the school? Is it intended to do all of these things? The Legislature and Mr. Stull have confused means with ends; they have prescribed the development of a tool but have not indicated with any degree of clarity what ends the tool is to be used for. Are, for example, evaluation procedures that are intended to improve the quality of teaching the same ones to be used to dismiss the grossly incompetent?

Given the context within which the section on teacher evaluation is found, teachers can respond in at least four ways. First, they can formulate or purchase behavioral objectives and then implement them through tests and other devices in order to demonstrate their effectiveness as teachers. Second, they can circumvent the Act by formulating objectives that they feel absolutely confident they can achieve and that are of a lower standard than those to which they would normally aspire. Third, they can ignore the Act by going through the motions to comply but not really taking it seriously. Fourth, teachers can as a professional body actively oppose the Act on educational grounds.

I predict that options two and three--circumvention and neglect--will be the most frequent methods for dealing with the Act. I believe this because I do not believe teachers will see the Act as something that serves their professional interests.

Dr. Popham points out in his paper, and I agree, that "by and large, the indicators of student progress . . . will be learner performance on measuring instruments." Gene Glass's paper, too, addressed itself to measurement. The Implementation Handbook of the Alameda County School Department also states that certificated employees' competence will be measured, but please note that the Act itself says nothing about measurement. It talks about assessment and evaluation. Why do we so quickly jump to the conclusion, as I believe many school administrators will do, that assessment and evaluation demand measurement? I do not have the time to spell out the difference between measurement and evaluation, but this confounding of the two simply reveals once more the conceptual confusion the Act has generated.

Dr. Popham's remarks about norm-referenced and criterion-referenced tests I applaud. For years, educators have employed achievement tests that have been designed to be irrelevant to educational purposes, since statistical criteria superseded educational ones in validating the tests. Furthermore, I applaud his remarks about the need for tests to be relevant to the curriculum. I worry about the people from school districts who come to my office looking for tests they can use back home. Tests which have no curricular validity have no place in the assessment of teachers.

Dr. Popham goes on to say that evaluation without objectives is meaningless. This assertion is also a result of a conceptual confusion that rests on the mistaken belief that the absence of objectives necessitates the absence of criteria for evaluation. This belief is simply false. We can and do evaluate an enormous amount of what we have learned from experience without first formulating objectives for those experiences. But Dr. Popham is right, I think, in believing that behavioral objectives will play a key role in districts bent on implementing the Act. I suppose all of us know how to write behavioral objectives. Behavioral objectives are considered by many to be the first, and indeed the crucial, step in curriculum planning and in evaluation. This belief is held in spite of the fact that the research literature on the effects of objectives on teacher planning and student learning is, even when generously interpreted, moderate or mildly negative about their supposed educational virtues. Studies by Baker, Oswald, Doty, Planey and McKie, Boardman, Bishop, Weinberg, Smith, Jenkins and Deno, Olson, Yelon and Schmidt, and a host of others have all failed to find educationally significant results from their use. But I'm afraid that, in spite of the empirical evidence, teachers will be expected to formulate such objectives to meet the provisions of the Act. Incidentally, to find out what professors of education at Stanford really think of behavioral objectives, I did a little study myself. I collected from faculty members teaching courses this quarter the handout materials describing the nature of and readings for their courses. The materials collected revealed that only one professor used anything that even remotely approached statements of behavioral objectives. Why is it we are so willing to prescribe for others medicine that we are unwilling to take ourselves?

Dr. Popham goes on in his paper, and I believe correctly, to point out that the appropriate level of specificity for objectives is itself a problem. The more one lays emphasis on the measurable, the greater the tendency towards specificity. In the 1920's, during the last gasp of the efficiency movement in education, literally hundreds of objectives were formulated by curriculum writers and by teachers for teaching each subject field. Such an enthusiasm can lead to a quagmire both for teachers and for those who are to evaluate them. Dr. Popham also points out that in setting standards, the entry behavior of the learners should be considered. Indeed, recognition of the differences among children in terms of what they have learned, how they learn, and how fast they learn has been the major motive for individualizing instruction. Do we want teachers to individualize instruction? And does individualization refer to the content and goals as well as to the rate and method of teaching? If so, how does this prescription for desirable educational practice square with the expectation that teachers formulate behavioral objectives for a class, or do we assume that a teacher should formulate behavioral objectives for each of his students? If so, how many measurable objectives is he to formulate for each student or for groups of students? Is the plea to individualize instruction--or, a softer phrase, to personalize teaching--compatible with the "uniform" evaluation requirements of the Stull Act?

Most of us here, certainly those of us who are university professors, have a tendency to view education from on high. The image of the ivory tower has a grain of truth in it. And when we look at things from on high, we tend, I believe, to oversimpilfy what we see. School administrators and state legislators suffer from the same problem. The planning of educational practice looks a lot simpler from a distance than from up close. It's conceptually neat to lay out goals and then evaluate effectiveness in relation to the goals as the Act provides. But at the same time that we embrace this procedure, we say we want teachers to be flexible and responsive to the interests of children. We say we would like them to engage in flexible purposing, to shift goals, to reorder priorities in light of the educational opportunities that arise in any classroom that is not hamstrung by rigid objectives and time schedules. Well, if we want these things, how do we reconcile this image of personalized, responsive teaching to prescriptions to predict where an entire class--even a set within a class--is expected to be by the end of the school year? How, in short, is a teacher to remain flexible with regard to goals, content, and method and at the same time to make long-range predictions of a behavioral variety in terms that are measurable? Perhaps it is this question that this conference is designed to answer.

The most egregious error that is made both by the Legislature and by educators regarding the Stull Act is the failure to recognize that the prescriptions in the Act are not merely prescriptions of a technical variety, they are value-saturated through and through and contain, albeit covertly, an implicit model of education. That model is, as I hope you are all aware, an industrial or production model of education. The tone of the Act is set in its opening passage. It's about the grounds (one of which is unconstitutional) for dismissing employees. The Act proceeds from that lofty height to a prescription to develop mechanisms that will make this possible. The production model is a comfortable one for lots of school administrators and even for some professors. It makes things appear simple; it provides a managerial approach to the control of the educational enterprise. The Act explicitly states that its intent is "the establishment of procedures and techniques for ascertaining that the certificated employee is maintaining proper control and is preserving a proper learning environment." Such a tone discloses as surely as descriptive language a view of what schools are to do and to be. But is the production model of education the only defensible conception of educational practice? Are there other ways of conceiving of teaching effectiveness than by formulating a "uniform system of evaluation" which in practice, I'm afraid, will technologize even more institutions that, in my view, would do well to increase their flexibility and playfulness.

Finally, what disappoints me the most is our failure to perform as a profession and as members of a university community one of our most important duties: to criticize the policies and practices mandated by government. It's too bad that our first thought is how to implement an act before we scrutinize the premises from which it springs and the practices it is likely to spawn. I think our communities deserve more.

In closing, I must add, lest anyone misunderstand, that I believe professionally responsible educational practice should employ sensitive evaluation procedures. I believe teachers should be accountable for what they do. But neither evaluation nor accountability must rest on prediction, on the overestimated usefulness of behavioral objectives, and on the use of measurement. There are alternatives.

What, then, do I suggest with respect to the implementation of the Stull Act? I suggest first that each of you who appreciate the implications and liabilities of the Act write to your state assemblyman to explain to him or her what these implications and liabilities are. In short, I suggest you act as educators. Second, I suggest that you return to your school districts and use the Act to help teachers talk about their purposes and to describe retrospectively what their students have learned. Third, I suggest that you explore with teachers ways in which critically descriptive appraisals can be made about the different sorts of things that individual students derive from their classes, and that you avoid generating competitiveness among both children and parents by comparing children with each other as the dominant, if not exclusive, mode of evaluation. Finally, I hope that when you talk with teachers you will leave them with the idea that you value diversity, experimentation, and risk-taking in the classroom, and through that encourage the type of educational flexibility that the four and one-half million children in California schools so badly need.

JOHN K. HEMPHILL

Far West Laboratory for Educational Research and Development

Dr. Eisner is a very hard person to follow--I am not sure that I can live up to the standards of his performance. I want to start out by noting first just what I heard Dr. Popham say. I find that his comments on the problems of implementing the Stull Act are essentially accurate and succinctly stated. It seems to me that these problems appear to him (and to me also) to be almost overwhelming. But he has made some suggestions that may possibly be useful to you who have the burden of getting the job done.

First, he notes that in evaluating the performance of teachers the Act is very explicit in prescribing that the evaluation be done by reference to what happens to the student. It is assumed that a good teacher is one who performs in such a manner that his or her students progress toward some established standard of excellence in each area of study. Although Jim points out some of the subtle difficulties inherent in this formulation of the task that are associated with what one means by standards or by progress, he refuses to be "hung up" by semantics or philosophic problems and moves to consider many of the practical issues of measuring "student progress." He correctly notes that the present state of the art of our ability to measure student progress may not be adequate for the job. He points out the strengths and weaknesses of norm-referenced testing and of criterion-referenced testing. He suggests that the implementers of the Stull Act have four options as they face major measurement problems, none of which is entirely satisfactory. First, an implementer may purchase norm-referenced tests from commercial vendors and risk their lack of relevance. Then later he may cope as best he can with the problems of interpreting what the results from such tests may mean. Second, he may develop his own tests. The tests are likely to be more relevant, especially if he can find the necessary resources in terms of both talented people and money, but in so doing he still risks the probability that because of limits upon his resources the tests will turn out to be less than entirely satisfactory. Third, if he is not willing to bear the technical burden of creating measures himself, he can look to a contractor who presumably has such expertise, and finally, he may join with others who find themselves in the same boat and thus share both the financial and technical burdens.

At this point Dr. Popham comes to the $64 question--what to measure--and faces it squarely. What are the objectives of teaching? That is, in what areas is student progress to be measured? He observes that this has not been a neglected problem in education. There are tens, hundreds, or thousands of statements of educational objectives--depending upon how thick or thin you want to slice them. Not everybody is in agreement about the relative importance of differently stated objectives, regardless of how they are sliced. Therefore, one must determine who is to make the selection of objectives and on what basis. I for one certainly agree with him that there is little need for each implementer of the Stull Act to start anew by writing his own list of objectives. Most will be much better advised to spend their efforts in selecting and ordering statements from the overabundance now stockpiled.

Finally, Dr. Popham comes to the problems of interpreting measurement data. Assuming that one can decide on an objective and can find or build suitable measurements which can be applied properly, what results can he hope to see and what meaning will they have for him? Is there some absolute standard of performance that, if reached, signifies effective teaching? How does one regard the input differences among classes of student with which the teacher has come in contact? that is, what do you do about the different levels of students and classes as you try to measure teachers' effectiveness? How can one interpret gain scores by testing from one time to another?

Surely Dr. Popham has laid out the problems. I am not sure, however, that he has really helped us much in suggesting what can be done about them. I am left with a distinct feeling that he regards the job as an impossible one and is ready to say why. I think he may well be right, and in fact I believe I could help him expand his analysis of why the Stull Act's mandates can't be complied with entirely. But to do so would be in no way helpful to you who, regardless of whether you can or not, must comply with the law. You may therefore rightfully ask me what I would recommend. Would I seek to have the Stull Act repealed? Would I ignore the whole thing and do something that would just serve as a smoke screen? Or, just what would I do? I think this conference is one answer to this question, and I hope it will prove valuable. By looking squarely at the problems we may avoid some of the obvious pitfalls and thus have some energies left to address the less obvious ones. I would recommend one other thing, and that is to try some "far-out" approaches with an experimental attitude. For example, why not constitute a small panel of lay citizens who could meet with a group of teachers and then render judgments of the progress students have made under a particular teacher. This could be done either in a highly structured manner or quite informally. Or the same approach, but with the judgments being made by teachers selected from another school, might be tried. As another example, teachers could provide documented evidence (themes, papers, etc.) for public inspection of their students' progress and the evaluators could look at a collection of these documents and judge whether progress is evident from year to year. One could collect these materials over a period of years and have in a folder the work a student has done in the fifth grade and have right beside it his sixth-grade work. This could include the best work the student had prepared, rather than a random sample. I think what i am saying is that any of you probably can come up with ideas about how you face the intent of the Act fairly and squarely and do some innovative, interesting, and useful things. I am not urging that we immediately rush to repeal the Stull Act. I think it will provide a needed stimulus and an opportunity for us to face the real problems, but I would recommend that we approach it with an experimenting attitude until we learn more.

EVALUATING NONTEACHING CERTIFICATED PERSONNEL

MAX G. ABBOTT
University of Oregon

As I read Assembly Bill No. 293 (the Stull Act) and the California State Board of Education Guidelines for School Districts to Use in Developing Procedures for Evaluating Certificated Personnel, two issues stand out as worthy of major emphasis in my presentation. First, the Act itself asserts that "it is the intent of the Legislature to establish a uniform system of evaluation and assessment of the performance of certificated personnel. . . ." (Note that I have put particular emphasis on the term "performance.") Second, the State Board guidelines stipulate that job descriptions, which should provide the basis for determining expected levels of performance for noninstructional personnel, should "include, insofar as possible, how and in what areas their services should influence pupil progress . . ."

The emphasis on performance as the basis for assessment is commendable. Moreover, the attempt to link the performance of administrators and supervisors to the performance of pupils directs attention to the ultimate reasons for the existence of administration and supervision in schools, which is to facilitate learning. This, then, is the good news. But now for the bad news.

When attempts are made to assess performance, and certainly when the assessment is conducted with the ultimate objective of improving that performance, it is essential to think in terms of effects. There is little point in attempting to alter the behavior of administrators and supervisors unless there is reason to think that in doing so some improvement will occur in the organization that will ultimately facilitate student learning.

The requirement to point to the effects of administrative acts, to establish causal relationships between the performance of administrators and the successful functioning of the school, confronts us with extremely serious problems at the present time. (In the remainder of this paper I shall use the term "administrator" to refer generally to noninstructional certificated personnel.) The fact is that there is very little dependable, vertified knowledge about such relationships. There is a good deal of writing on the topic that is speculative, some of it even persuasive; there are bits and pieces of empirical research available that are suggestive; and there is a limited body of theoretical literature that posits hypotheses and that lends credence to some of the speculations. But definitive, long-range research remains to be done on the subject.

We cannot wait until the research has been completed, however. The California legislation requires action now, a requirement that could produce beneficial side effects. The mere attempt to apply what is now known will serve to dramatize the fact that much more needs to be learned. In the meantime, efforts must go forward to identify the most plausible sources of the meaningful effects of administrative performance and to use those sources as a basis for assessing (and improving) that performance.

We can begin by clarifying the difference between the role of the administrator and the roles of other functionaries in the school. Whereas teachers, counselors, nurses, and psychologists work directly with students and thus are intended to affect them directly, administrative performance is only indirectly related to student behavior and achievement. The direct output of administrative performance is improvement in the environment in which teaching and learning occur. Thus, it is the effects of the administrator's performance upon teachers and other functionaries as a group --upon the organization--that provide the basis for determining his or her effectiveness. If the administrator's performance enhances the effectiveness of the faculty, both individually and collectively, he is an effective administrator. If his performance inhibits the effectiveness of the faculty, either individually or collectively, he is an ineffective administrator.

To get at the question of administrative performance, then, it is useful to examine those functions of administration that contribute in essential ways to improving organizational conditions. There are three such functions that appear to be paramount: decision making, conflict management, and organizing.

Decision Making

To say that decision making is the essence of organizational functioning is merely to state the obvious. It follows, then, that effectiveness in facilitating the making of decisions looms large in any set of criteria for assessing administrative performance. But what are the particular decision-making skills from which such criteria might be derived? I would suggest five: (1) skill in differentiating among types of decisions; (2) skill in determining the amount and type of information needed to reach decisions; (3) skill in determining the appropriate involvement of other people in reaching decisions; (4) skill in establishing priorities for action; and (5) skill in anticipating both intended and unintended consequences of decisions.

Differentiating among Types of Decisions

Much of the literature on decision making reads as if all organizational decisions were of a kind and thus could all be approached using the same procedures. It is obvious, however, that this is not the case. Decision situations vary in a number of dimensions, resulting in wide variations among the types of decisions that must be made.

A number of classification schemes might be used to indicate the differences that exist among the decisions that are made in an organizational setting. For example, decisions could be differentiated in terms of their implementation requirements and the implications those requirements hold for participation in the decision-making process. If a decision can be implemented only if it is fully understood and supported by a school faculty, obviously the faculty needs to be involved in making the decision. On the other hand, if a decision can be implemented merely through administrative action, and particularly if the content of the decision is of little consequence to faculty members, then the involvement of the faculty not only wastes valuable time but is also damaging to faculty morale.

Regardless of how one classifies types of decisions, it is obvious that differences do exist among decisions that must be made in a school, and that those differences mean that there are also differences that must be taken into account regarding the amount and type of information needed to support a decision, the procedures that should be used to reach a decision, the individuals who should be involved in decision making, and the steps that need to be taken to implement the decision once it is made. The ability of an administrator to discriminate among issues and thus to differentiate among types of decisions is essential if an organization such as the school is to take appropriate action.

Determining the Information Needed for Decision Making

A number of people have called attention to the importance of information in decision making, pointing out that the search for information constitutes a significant first step in the preparation for making a decision. The search for information is, in fact, more pervasive than that; every step in decision making requires information--information concerning the nature of the problem, information concerning solution requirements and alternatives, information regarding the probable consequences of adopting various alternative solutions, information about the feasibility of different approaches to implementing a decision, and information regarding effective performance once a decision has been implemented.

The administrator who consistently makes decisions without adequate information will frequently find himself enmeshed in problems of his own making. This may occur when he initiates action to solve a problem that does not in fact exist, when he fails to recognize a problem that is literally clamoring for attention, or when his perception and definition of a problem are so inaccurate that any action taken is destined to be inappropriate.

Just as it is possible to err by taking action without adequate information, however, so is it possible to err by delaying action unduly while a search for further information is undertaken. Those who would postpone making a decision until _all_ of the information is at hand need to be reminded that _all_ of the information can never be assembled. Problem situations do not remain static, but change with the passage of time. It is never possible to know whether one has reached the best decision; it is only possible to determine whether one has reached a satisfactory decision. One aspect of effective administrative performance, therefore, is the ability to determine when sufficient information has been assembled to justify action.

Determining the Appropriate Involvement of Other People

Much of the educational administration literature has been confusing on the issue of employee participation in decision making. The writing on "democratic administration" has been particularly misleading in this respect, frequently conveying the impression that all employees have an inherent right to be involved in every decision that is made.

There are obviously a number of reasons why employees should participate in the making of some decisions; indeed there are many occasions when such participation is essential. As I suggested above, a decision that depends upon the full understanding and support of a faculty for its successful implementation requires the participation of that faculty. Other decisions require such participation as a means of assembling adequate information, regarding either the nature of the problem or the probable consequences of alternative courses of action. There are a number of additional reasons why employee participation is frequently advantageous for the organization, including the fact that such participation tends to decrease the visibility of the differential distribution of power in the organization and thus tends to enhance employee morale.

The fact remains, however, that there are occasions when decisions must be made unilaterally, without the direct involvement of employees. Sometimes this occurs because the urgency for action is such that there simply is not time to allow involvement to occur. At other times, the decision to be made is of little direct interest to employees; in such cases their involvement would be viewed by the employees as a waste of time on their part and as an indication of timidity on the part of the administration. Finally, there are some decisions that are of such overriding importance to the board or the administration that they must be made from a managerial perspective, even at the risk of creating internal dissension and of lowering employee morale.

An effective administrator must be able to make defensible judgments in this area. He needs to be skillful in recognizing when involvement is needed and when it is unnecessary, in judging when participation is essential for implementation and when it is irrelevant, and in determining whether an issue is in fact of unusual importance to the management of the organization or whether such a determination is used as a rationalization to avoid sharing decision-making prerogatives.

Establishing Priorities for Action

One of the first facts of life that persons new to the administrative position discover is that they will have more problems and issues brought to their attention than they can possibly deal with. Some will be crucially important, many will be trivial; some will require immediate attention, others can or should be deferred; some will require the personal attention of the administrator, others can be delegated. In other words, any administrator is frequently, if not constantly, confronted with the necessity to establish priorities for action. He can do so deliberately and with forethought or he can allow priorities to be set fortuitously, by default. In either case priorities will be established.

The administrator who moves deliberately to establish priorities is performing in a proactive mode. He maintains some control over the way he spends his time and he increases the probabilities that he will remain "on top" of problems. On the other hand, the administrator who allows priorities to be established by default is performing in a reactive mode. He will frequently be engaged in putting out brush fires. Many manageable problems will be allowed by him to become unmanageable.

The skillful setting of priorities, however, involves a great deal more than merely reviewing a series of items and establishing a time schedule for action. Priority setting requires the judicious use of information, some of which may be readily available and some of which will need to be assembled, and it involves skill in anticipating the consequences of acting immediately or of deferring action, of taking action personally or of delegating action to others.

Anticipating Consequences

Decision making consists essentially of estimating the outcome states of a system. That is, reaching a decision involves making choices based on predictions of the probable consequences of alternative courses of action. It is characteristic of all decision making, however, that no decision produces only one effect or set of effects. Moreover, the consequences of any decision are never limited only to those that are intended. All decisions result in multiple consequences, some intended and some unintended, some anticipated and some unanticipated. A decision to initiate a course of action to solve one problem may create other problems. It may also generate an entirely different response from the one intended.

An accurate assessment of the probable consequences of a proposed solution strategy (decision) will greatly reduce the chances that an organization will be taken by surprise. In some instances, such an assessment will reveal probable unintended consequences with potential for serious disruption. With advance information, however, the organization is able to take steps to minimize or to counteract those disruptive effects. In other instances, harmful aspects of the unintended consequences of a proposed decision will be judged to be so seriously inimical to the organization as to render that decision untenable. Such judgments obviously cannot be made unless those unintended consequences are anticipated. A crucially important administrative skill, therefore, is the skill of elaborating and making explicit the possible consequences, both intended and unintended, of each solution strategy proposed, and thus of reducing to a minimum the unanticipated consequences of each solution strategy adopted.

Conflict and Conflict Management

Whenever two or more people join together in purposeful activity, the potential for conflict exists. Although conflict is not necessarily dysfunctional for an organization--indeed there is a substantial body of literature which deals with the positive functions of organizational conflict--the potential for disruption and dysfunction is always present in a conflict situation. Whether a conflict situation turns out to be functional or dysfunctional will depend frequently, often centrally, on the accurate analysis of the forces involved and on the response that is made to those forces.

To be effective, therefore, a school administrator must develop the ability to analyze and diagnose conflict situations and to choose courses of action that respond both accurately and adequately to those situations. An inappropriate response to an emerging conflict will frequently exacerbate that conflict; at the least, it will do nothing to reduce or ameliorate the conflict.

Among the many conceptualizations that exist in the literature on organizational conflict, one that is particularly useful for present purposes is the one that has been developed by March and Simon[*] and that utilizes the source of the conflict as the basis for analysis. In this conceptualization, conflict is viewed as arising when either of two conditions prevails. First, conflict occurs when there is disagreement regarding the goals that are to be pursued or the values that are to be implemented. Second, conflict occurs when there is disagreement regarding appropriate actions to be taken in pursuing goals or implementing values, even though there is agreement about the goals to be pursued or the values to be implemented.

Where there is agreement on goals and values but disagreement regarding appropriate actions, the typical response to conflict follows the classical problem-solving pattern. That is, a search is generated to obtain information about possible alternatives, attempts are made to predict the consequences of the various alternatives, and efforts are initiated to calculate the costs and benefits of each alternative. The response to this type of conflict situation, then, is based essentially on logic and rationality. The procedures used to resolve the conflict--gathering, interpreting, and assessing information--place minimal strains on organizational relationships, and the decisions or solution strategies adopted generally are designed to maximize gains and to minimize losses for all concerned.

When there is fundamental disagreement over goals and values, on the other hand, information regarding alternative courses of action and probable outcomes is largely irrelevant. Parties to the conflict, therefore, tend to view information gathering as a meaningless activity and a waste of time. The typical response in such a situation is to attempt to bargain or to marshal sufficient power to "win." These procedures place considerable strain on organizational relationships, particularly those with an authority dimension. The decisions or solution strategies adopted almost universally involve a win/lose situation; for one party to gain another party must lose.

---

[*] J. G. March and H. A. Simon, Organizations (New York: John Wiley & Sons, 1959), p. 141.

Because bargaining and "political action" do place strains on organizational relationships, and since those in authority positions feel particularly vulnerable in a win/lose situation, there is a strong tendency for organizations, and particularly for administrators in those organizations, to treat all conflict as if it were based on disagreements over procedures and thus amenable to resolution through problem solving. In the case of value conflict, such a response leads to the use of organizational resources for activities that are both irrelevant and futile. Meanwhile, the conflict persists and exacts its own toll in organizational energy. Thus the inappropriate response not only fails to use the conflict to move the organization ahead, but it frequently intensifies the conflict and seriously reduces the organization's effectiveness.

Although it is sometimes possible to deal with a conflict involving disagreement over goals and values using a problem-solving approach, this cannot be done if there is a failure to recognize or to admit the nature and extent of the disagreements. It can only be accomplished by recognizing and admitting that such disagreements exist and by moving to a new level of possible agreement.

Not all conflict situations can be handled this way, however. Some cannot, nor should they be. Bargaining and "political action" are thoroughly legitimate approaches to conflict resolution. Discomforting though they may be to the hierarchy, under certain conditions they represent the most appropriate, and sometimes the only, approach available. And, employed judiciously, they can be a means of moving the organization to a new level of functioning.

To use conflict, either latent or manifest, to the advantage of the organization, however, it is essential that administrators have the ability to analyze and understand the situation, select a strategy that is both responsive and appropriate, and implement that strategy intelligently.

Organization and Management

A central task of administrators is to take the lead in establishing and maintaining an effective organization for carrying out the functions for which the institution exists. This involves two major classes of action: (1) establishing structural arrangements that are appropriate for the functions to be performed; and (2) providing adequate incentives to elicit the contributions from members of the organization that are necessary for the accomplishment of the organization's tasks.

The administrator has a major responsibility for assessing the tasks to be performed in accomplishing the organization's purposes and for deploying resources (human and nonhuman) appropriately according to those tasks. This responsibility involves the selection and assignment of personnel; the development of job descriptions and requirements; the procurement, allocation, and utilization of facilities and supplies; and the establishment of schedules for the effective and efficient use of time.

The way a unit is organized to perform a given task or set of tasks should be determined by the nature of the tasks, whether those tasks represent temporary responses to ad hoc issues or whether they consist of programmed activities designed to carry out essential and enduring organizational functions. The establishment of a committee, for example, could be an appropriate way of organizing to perform an ad hoc task if the task calls for obtaining a variety of points of view and considering a number of alternative courses of action. There are other tasks, however, such as the writing of a document, that represent primarily the efforts of an individual and for which a committee may be wholly inappropriate.

In a similar vein, the use of modular scheduling, the adoption of team teaching, or the development of other arrangements for the use of time, space, and personnel for instructional purposes can produce beneficial effects. To do so, however, it is essential that all those participating understand the nature of the structural changes and the demands that they place on the interaction patterns of both staff and students. They must also recognize the necessity for altering approaches to instruction and the use of learning materials that arises when new structures are adopted. Such understanding and recognition are not likely to occur unless the administrator is able to conceptualize the relationship between structure and function and can thus take the lead in making the necessary accommodations.

Although the development of structural arrangements and the procurement and allocation of resources are important steps in organizing for work, these tasks represent only a first step. An effective organization is one in which the members, usually employees, understand and appreciate the contributions that they and others make to the overall purposes of the institution and in which a personal sense of achievement accompanies the performance of institutional functions.

This means that the administrator must constantly be aware of two imperatives for organization functioning. First, there must be visible evidence of movement in the sense that the purposes for which the organization exists are being accomplished. Whether this is referred to as task organization, thrust, production emphasis, or some other term denoting accomplishment, it is an essential element of organizational functioning. Second, the members of the organization must be provided with rewards or incentives that are commensurate with the contributions they are expected to make. The most obvious rewards are those that take the form of salaries and other monetary benefits. But there is massive evidence to demonstrate that monetary rewards alone are insufficient inducements to sustain wholehearted efforts over a period of time. A sense of personal achievement, the recognition of others, opportunities for advancement and growth--all these must accompany work if that work is to have meaning in an individual's life, meaning that is essential to sustain morale and the motivation to continue to contribute.

The actions of administrators determine to a large extent whether these two imperatives are fulfilled. Some administrators become so preoccupied with task accomplishment that they lose sight of the fact that they are dealing with human beings who have their own aspirations for achievement and their own needs for personal fulfillment. Other administrators become so thoroughly imbued with the human element that they forget the organization exists for purposes beyond the happiness and comfort of the employees. To understand the interaction of organizational demands and human needs, of task accomplishment and personal fulfillment, and to perform accordingly, is an important aspect of effective performance in the administrative role.

Although there is yet a great deal to be learned about the effects of the performance of administrators on other members of the organization, there are some indicators that can be used to determine whether an administrator functions effectively in this domain. One such indicator can be found in the way an administrator uses information and formulates problems. Competent human beings tend to view problems as a challenge; they respond to problem situations by attempting to effect a satisfactory resolution to the problems confronting them. This happens, however, only if the problems are seen as real and of direct concern to those who must participate in solving them. An effective administrator, therefore, is alert to information that enables him to sense that a problem either exists or is arising, and he seeks ways to share that information with others so that they see the problem as one calling for action.

The incentive to perform is also affected greatly by the ways administrators use rewards and sanctions. Although teachers, like employees in other organizations, look to a variety of sources for social support, they are particularly sensitive to the way administrators respond to their performance. If their attempts to be creative and imaginative are met with indifference or hostility on the part of administrators, then the incentive to be creative and imaginative will be seriously curtailed. If recognition, including promotion, consistently goes to those teachers who are submissive and passive, then submissiveness and passivity become the accepted norms for behavior.

All of this is a way of saying that administrators set the tone for the way an organization functions. Another way of viewing organizational tone, or climate, is through an analysis of organizational morale, with particular emphasis upon employee satisfaction.

In a number of studies in both industrial and educational settings, a distinction has been made between the factors that contribute to employee satisfaction and those that create dissatisfaction. Consistently, the factors that have been shown to contribute to satisfaction are: a sense of achievement; recognition of others; opportunities to accept responsibility, both for one's own efforts and for the work of others; and opportunities for advancement and for professional growth. Just as consistently, the factors that have been shown to create dissatisfaction are: unfair policies and administrative practices; incompetent supervisors; lack of recognition of others; and poor interpersonal relationships with supervisors and administrators.

Two implications of these studies are worth noting here. First, the factors that contribute to dissatisfaction tend to center around the administrator and the forces that he controls. Second, unless the conditions that create dissatisfaction are corrected, the factors that create satisfaction tend to be submerged. Thus, it appears that the administrator has the potential both for contributing to the dissatisfaction of employees and for blocking the emergence of those factors from which employees obtain a sense of satisfaction and fulfillment in their work.

The means of obtaining a reasonable measure of employee satisfaction are relatively simple and inexpensive to administer. Thus, the organizational conditions affecting satisfaction and dissatisfaction provide another source of meaningful criteria for effective administrative performance.

DISCUSSION OF PAPER BY MAX G. ABBOTT

LEWIS ALLBEE
Barstow (California) Unified School District

Before I give my reactions, I would like to preface them by saying I am not speaking from an "ivory tower." I'm on the firing line along with many of you and probably see things much as you do. Second, I would like to say that, since the Stull Act is now the law, we are expected to follow through with it. The citizens of our community, the parents, the school board members, and others expect us to do something, and we are trying to do it.

From the point of view of a practical administrator, I liked very much what Max Abbott had to say and I subscribe to his overall analysis of the administrative functions. It is also my understanding that the Stull Act requires us to identify the meaningful effects of administrative performance as they relate to effective teaching and learning in the classrooms.

His analysis of the three major aspects of administrative performance as being decision making, conflict management, and organizing helped to pinpoint those areas that need to be looked at as sources of criteria for evaluation. The sources of criteria that he suggested for evaluating performance in these three areas are very useful: for decision making he included at least five separate skills; for conflict management, three separate clues; and for organization and management, three more valuable suggestions. I think perhaps the idea that struck me most deeply was his statement that administrators "need to understand the interaction of organizational demands and human needs, of task fulfillment and personal fulfillment, and to perform accordingly." The indicators that he gave revealed ways in which we may help promote a real understanding not only of the idea of the production we want but also of the human factors that are involved. He mentioned that the factors that contribute to dissatisfaction tend to center around the administrator and the forces he controls. I think that is generally true; however, many of us find ourselves at the present time with shrinking enrollments, over which we haven't much control. Shrinking enrollments usually mean a limitation on income, and the moment there are limitations on income, the administration is not free to do the things that it might like to do.

Finally, it was stated or implied that the administration should provide the members of the organization with awards or incentives that are commensurate with the contributions they are expected to make. I certainly agree with that, but I'm not sure that we are able to do it nor that we ever will be able to do it.

In closing, I would like to say that, bad as the Stull Act may or may not be, or bad as the mechanical arrangements may be that we are trying to develop, we are nonetheless trying to come up with something that will work, and I think we will be successful.

JAMES A. KLAWITTER
California School Board Association

My point of view is also that of someone working in the field. The term accountability is relatively new--the need to be accountable is not. Critics of education, both friend and foe, tell us that typical management practices in education are in many ways inhumane. One writer likens the total educational bureaucracy to a big marshmallow. External pressures make little depressions here and little displacements there, but when the pressure is relaxed the marshmallow quickly restores itself to its original form; it is very resilient and very self-protective. Now most Californians believe or want to believe that their schools are doing a good job, but they are hungry for definitive, visible evidence of institutional effectiveness and their hunger is expressing itself in accountability language. This in turn creates a dilemma for many administrators. It has become fashionable to shrink from making judgments; thus systems are devised or machines or statistics are devised to make the judgments. Institutionalisms of judgments or depersonalized appraisals are employed.

On the other hand, some writers tell us that administrators are cautious in regard to accountability, not because administrators are irresponsible or complacent but because there are insurmountable obstacles to the evaluation of administrators. At any rate, to do or not to do so has been settled for us by law. California law commands administrators to be accountable and that's where we find ourselves.

In this connection, Dr. Abbott refers to the lack of valid data to support a causal relationship between the performance of administrators and the successful functioning of the school, and within that limitation he delineates a matrix of skills, responses, imperatives, indicators, and factors of effective administrator evaluation. As he mentioned, these items can be used as the basis for an evaluation of administrative performance.

I believe Dr. Abbott's paper is an important contribution to the deliberations of this conference. My first comment on it deals with the topic of conflict resolution. This function is very applicable to the current California scene, where administrators increasingly find that their meet-and-confer efforts end in persistent disagreement (as defined in the Educational Code, 1387.1). Dr. Abbott points out that where conflicts involve a fundamental disagreement over goals, information gathering is a waste of time and the typical response is the use of power by one or both parties. He says that the solution is for both parties to recognize the nature and extent of the disagreement and to move to a new level of possible agreement or functioning. He also proposes the use of bargaining or political action as alternatives. I suggest that this is a difficult situation for administrators. I know few administrators who want or like to use conflict in district management. Some few can skillfully use conflict to the advantage of the school district's organization. Conflicts have enormous impact on district personnel and on programs, and they require great skill in their resolution. The task of evaluating the performance of an administrator involved in conflict resolution requires special consideration and new skills and processes that I don't think we possess at this time. It is a task that schools and administrators cannot ignore, however, because conflict--whether spontaneous or calculated--is becoming, and perhaps has already become, a way of life.

My second comment is to suggest consideration of an additional dimension relative to the function of the educational administrator. I refer to the function of knowing and satisfying the needs of the public. The tensions between educational professionals, including administrators, and community interest is growing. Spokesmen for the public conscience carefully watch the schools. The effective educational administrator learns that educational accountability is a mixture of social problems, political pressures, and economic forces, in addition to educators' expertise and formal educational experiences. A solution to the accountability conflict between the public and the schools organization might be for the public, the government, and educators to admit that there are learning and management problems in the schools and that none of the three interests knows how to solve all of them, none can guarantee maximum educational achievement for all, and participatory efforts are needed from all parties.

Administrators need to recognize that accountability as simply a professional exercise is unworkable. It must be combined with community effort, where at best the politics of accountability is an exceedingly tricky domain. There is a reference in Arnold Tennenbaum's work, Control in Organizations, in this regard. He says that an infinite amount of control exists in any organization and that subordinates can increase their control without the super-ordinate having to relinquish any of his control over the organizational decision-making apparatus. Tennenbaum also found that when the total amount of control was high, indicating much interaction, there was also a higher degree of commitment to the organization, and that such organizations tended to be more effective. Now whether these conclusions are valid, or valid in all instances, or whether administrators will act or will be allowed to act on these premises, I don't know. But it is clear that accountability in education must be the result of rational understanding and communication between public and educators, including administrators, in regard to determining educational purpose, defining function, judging results, and taking corrective action to improve learning.

This brings me to some comments on the area that is the meat of Dr. Abbott's paper--the functions of the educational administrator, namely decision making, conflict management, and organizational management. It is apparent that a major step is needed to mold better school organizations, and that is to recast the role of the administrator. Some recasting has taken place already in some school districts as a result of the meet-and-confer operation. This has been limited primarily to instructional personnel. Other factors have been responsible for a limited recasting of the administrator's role, but we are still talking of the emerging role of the principal, the emerging role of personnel directors, and so on. I suggest that this "emerging role" language is mostly nostalgic crying over spilt milk. An update of management practices is long overdue. For example, the veto power, through passive resistance, of professional administrators (especially in large districts) is virtually unbeatable. David Rogers in his book, Politics and Bureaucracy in the New York City Schools, speaking of the New York City schools, says that the school district has almost unlimited capacity for absorbing protest and externalizing the blame, for confusing and dividing the opposition, for seeming to appear responsive to legitimate protest by issuing sophisticated and progressive policy statements that are poorly implemented (if at all), and then for pointing to all the paper accomplishments over the years as evidence of both good faith and effective performance.

I believe one of the roots of this problem is the lack of theoretical basis for much of educational administrative thinking. Fortunately, much of the newer educational writing is based upon research findings. Educational administrators want to see the evidence. I suggest that some of the administrative functions referred to in Dr. Abbott's paper could be realized and beginnings could be made on administrator evaluation if there could be greater cooperation between educational administrative practitioners and educational administrative theorists. I suggest also that we must settle for probable rather than absolute realities at the moment in developing administrative evaluation systems.

In conclusion, some of the problems that I see in evaluating educational administrators include the subdued nature of the accountability ethic within the fraternity of administrators, the lack of effective bargaining and conferring on administrative actions, the lack of knowledge in regard to relationships between educational input and educational output, the lack of a valid basis for discovery and improvement of good practices in education, and the need to understand institutional change in achieving such transformations. We need to achieve equity of educational opportunity at all levels. We also need to correlate the level of quality among programs with particular levels of expenditure. We need to determine the accountability expectations of the many public interest groups and educators who comprise the school district. We need to assign a higher priority to accountability and administrator evaluation on the list of educational needs. And we need to avoid the situation in which the long-term effect of the measurement and evaluation of administrators will be a trend to standardization and patterns of uniformity, as compared with individuality and diversity. I'm sure all of you see many other problems.

DUNCAN SPRAGUE
California Teachers Association

It is not just teachers who are involved in the Stull Act--indeed, the Act does not even say "teachers," it says "certificated employees." All morning I have been hearing how teachers are going to be evaluated. Then I listened to Dr. Abbott talk about the factors that contribute to dissatisfaction--unfair practices, the fact that interpersonal relationships are not good, etc.--and turning to Section 5.5 of the Stull Act I found there is no differentiation between teachers and nonteaching certificated personnel except in Section 13486, which states that the governing board shall avail itself of the advice of the certificated instructional personnel in the district's organization; in all other places we find "certificated personnel" rather than "teachers," as stated in the State Board guidelines.

I was thinking I might start these remarks with a disclaimer--I'm on the front line too. I go around to school districts and have an opportunity to talk with staffs, and I find that many of them have not read the Act. As I go throughout the state, and listen, and read, I find that many attempts to implement the Act are simply taking out the old boilerplate, polishing it up, and calling it a new industrial accomplishment. And it really is not.

The other day I read a principal's objective for the second grade. He stated that all students in the second grade shall read at the second-grade level by the end of the year--this is his sole objective for the second grade! I wonder how he is going to accomplish that--with a whip and a chair? He did not state his duties, or that he was going to try to support the teacher to make sure that materials got there and that she didn't get one reading book for every five children. I think when we start talking about implementation of the Act, much of the sound judgment and proposals that Dr. Abbott has set forth regarding decision making must come into effect.

Let us assume we receive a proposal headlined "Implementation of the Stull Act." We turn the page and it says "teacher evaluation"--and that is all it applies to, teachers. But to implement the Stull Act requires more than teacher evaluation, because the Act says all certificated employees shall be evaluated and assessed. Also, many of the evaluation systems being adopted state, in effect, that a teacher is to diagnose every single person in his or her class and prepare behavioral objectives by October 1. I don't think the Mayo Clinic could do a diagnostic job on one individual as rapidly as many of our teachers are asked to diagnose individual learning difficulties and the learning situation in an entire classroom. I dare say, to the concerned administrators here, that this is the type of situation that you don't want to occur.

Not only do we have the problem of having diagnoses complete by October 1, but we are told that the teacher is going to be evaluated at the end of the year on the behavior objectives, on a pretest-posttest basis. I normally ask the administrator at that point to read Education Code Section 13488, and I ask how he is going to fulfill its requirement if the teacher is to be judged on behavioral objectives at the end of the year. In case you don't have a copy of the Act, it states that the evaluation assessment made pursuant to this article shall be reduced to writing and a copy thereof shall be transmitted to the certificated employee not later than 60 days before the end of the school year. But I wonder, how are we going to judge how well a teacher has done at the end of the school year and write a written evaluation of the teacher 60 days prior to that?

I am point out some of these things because I was inspired by what Dr. Abbott had to say about decision making. What I see in his remarks is that you gather information before you make decisions, and I'm trying to give you some input on some of the difficulties we see throughout the state.

I would like in conclusion to repeat something Dr. Abbott said early in his presentation. The direct output of administrative performance is improvement in the environment in which teaching and learning occur. I think, if this is the direct output that comes as a result of the decisions made after tomorrow, then in truth we will do something for the public school students in California.

THE ASSESSMENT OF PROPER CONTROL AND SUITABLE LEARNING ENVIRONMENT

NED A. FLANDERS
Far West Laboratory for Educational Research and Development

and

GRETA MORINE
California State University, Hayward

A Point of Departure

This paper is concerned with criteria which can be used to evaluate how supervisors, principals, and teachers perform their adjunctive duties, help to create proper control, and jointly produce a suitable learning environment for the young people who are required by law to attend school. Criteria of evaluation refer to things that can be measured, so we must first ask what measurement techniques are valid and feasible.

If the goal of assessment is a one-time evaluation to be used for purposes of determining retention and merit pay increases, then, for assessment to be valid, we need:

1. Strong general agreement on the meaning of proper in "proper control" and suitable in "suitable learning environment"
2. Strong research evidence linking particular teacher behaviors to both control and the learning environment
3. Measurement instruments that are well developed, that measure most if not all of the facets of behavior related to control and learning environment, and that probably will be administered by outside consultants rather than by local school personnel, in order to maintain objectivity of results

It is not feasible to satisfy any of these three requirements, given our multivalued culture, the current state of educational research, and the present financial situation of most California school districts. Therefore, assessment viewed as a one-time evaluation cannot be valid and feasible at the present time. It is somewhat ironic that the legislative committee hearings on the Stull Act apparently provided no testimony (McDonald, 1972) on the adequacy or inadequacy of the currently available evaluation techniques. However, since the Stull Act makes provisions for counseling and advising certificated personnel after an evaluation, we can assume that there is a strong emphasis on the improvement of performance.

When the goal of assessment is seen as learning how to improve instruction, a valid assessment requires:

1. Agreement between the teacher and the supervisor on the instructional goals and/or the instructional theory to be utilized
2. Some research evidence linking particular teacher behaviors to control and learning environment
3. Some measurement instruments that can be administered by local school personnel as well as by outside consultants
4. Some training materials to enable teachers and administrators to improve skills when a need for improvement has been identified

There are areas of performance for which these requirements can be met. Therefore, assessment aimed at improvement of instruction can be considered to be both valid and feasible. Feasibility will be increased if at least some of the available training materials are self-instructional. This will make skill improvement more individualized and less expensive.

Clarifying Terms

In order to discuss how the provisions of the Act dealing with adjunctive duties, proper control, and suitable learning environment can be implemented, it will be necessary to propose a meaning for the various terms that appear in this paper. The two terms which stand out as begging for clarification are *proper* control and *suitable* learning environment. A good deal of space could be assigned to this task, but we prefer to take a shorter, pragmatic approach. The definitions below were determined largely by the limited number of assessment techniques that are available and the mandate of the Act, which is, to start *now*. Those readers who would prefer to define these terms differently should remember to reconcile their preferences with the operational meanings that the assessment techniques provide.

1. The *goal* of evaluation and of the Stull Act itself is the *improvement of education*, particularly teaching.

2. *Proper control* ultimately rests on a judgment that the methods of instruction are consistent with the purposes of instruction. Most teachers work in school districts which have a well-established curriculum of study. Yet any curriculum is primarily a guide that a teacher constantly adapts to his instructional purposes, his style of teaching, and his students. When this process of adaptation results in all of the students using the same materials and working on the same assignments under relatively close teacher supervision, a more formal and traditional pattern of active teacher control will seem consistent. When a teacher chooses to create the freedom necessary for independent, individualized learning activities, then a more flexible pattern of control will seem consistent. All teachers exert control. We can define proper control as the *teacher's ability to create and maintain learning activities in which the intended learning can occur without undue*

disruption. The judgment of what is proper and what is improper will not be the same for different adaptations of the curriculum. For example, freedom to talk to another student, move about the room, choose one's own learning task, etc., might be considered a symptom of improper control in a formal classroom but such freedom would be proper in a more open classroom. We might further note that proper control is essentially an adult concept. The values of the parents, the teacher, the principal, or those who provide professional supervision are usually the primary determinants of what is "proper."

3. A suitable learning environment, on the other hand, is essentially a child-centered concept. A learning environment refers to the students' opportunity to exploit a situation for the purpose of learning. A suitable environment means that the arrangements for learning match the interests and abilities of the students who are there to learn. One important aspect of a suitable learning environment is that students can participate effectively no matter what their ethnic origins may be and regardless of the socioeconomic level of the community. There are many features of a learning environment that we could emphasize, but for it to be suitable, we would at least expect the students to be interested in coming to class, to look forward optimistically to the work involved, and to obtain a sense of satisfaction from participating, especially in terms of self-respect and self-confidence.

4. Adjunctive duties are presumably to be distinguished from active teaching, administering, or supervising. One might guess they were added to the Stull Act by those who could remember when teachers also taught Sunday School, principals were expected to plan the Community Chest drive, and supervisors were expected to join a hospital auxiliary. In today's schools adjunctive duties may involve educators in projects to increase the parents' participation in the school.

to revise curricula, or to redesign report cards. The most important adjunctive duty related to classroom control and learning environment is taking the responsibility for analyzing one's own performance, exploring professional alternatives, and deciding when an alternative is an improvement; to this we add assisting others to do the same.

5. Competence and standards are probably used in the language of the Act in the simple sense of performing one's job in a satisfactory manner. In this paper we point out that information about an individual's performance can be used in several ways. To invoke standards usually means that criteria of performance can be identified and assessed, that certain minimal levels can be established, and that individual performance measures can be compared with norms from an appropriate reference group. Unfortunately, it is difficult to invoke standards when adjunctive duties are not consistent from one individual to the next, when proper control varies from one situation to the next, and when a learning environment becomes suitable precisely because it is custom-built to meet the needs of a particular group of youngsters. Faced with this diversity, we suggest that information about performance can be compared with behavioral descriptors of selected skills which, in turn, might be derived from a model of teaching. Another alternative is to compare an individual's performance at one point in time with the same performance at another point in time in order to determine the presence or absence of change. These applications of "standards" are likely to improve instruction provided those in the field act wisely by adapting the provisions of the Stull Act and using research results judiciously.

6. Student progress is dealt with in another paper at this conference, but it will be discussed here in terms of outcomes other than subject matter achievement. Attitude inventories can be used to assess student perceptions of the

teacher and the learning activities. Such measures are associated with patterns of verbal interaction (Flanders, 1970, p. 389). We would define student progress to include these student perceptions of the learning environment as well as patterns of desirable student behavior that can be observed during learning.

Adjunctive Duties

Both teaching and nonteaching certificated personnel now face the task of evaluating professional performance. We have defined this task as an adjunctive duty which has been added to all certificated job descriptions. As a result, one can now say that it is the duty of supervisors, principals, and teachers to participate in the evaluation of their own professional performance and to help evaluate the performance of their colleagues.

The same steps of overall evaluation can be used with a wide variety of adjunctive duties such as projects to increase parent participation in the schools, to redesign report cards, or to change the curriculum.

Procedure	Features to be Evaluated
1. The purpose of the project is decided by conferring with those who will participate.	Is the project easy/difficult, significant/insignificant, responsive to current priorities/not responsive, etc.
2. Plans are made to assess progress at appropriate check points.	Are plans practical (space/time/resources); are objectives specific/ambiguous, consistent with purpose/inconsistent, etc.

Procedure	Features to be Evaluated
3. Project is carried out.	Were final objectives and check point objectives achieved/not achieved; did a change occur/not occur; can the nature of change be determined/not determined, etc.
4. The consequences of the project are judged to be an improvement or not an improvement.	Were there sufficient/insufficient data to determine inprovement; can results become part of the program/not part of the program, etc.

This outline is a practical plan of evaluation which would have been appropriate at any point in the history of public education. Given the progress that has been made in the 1960's in analyzing teaching, this old outline can now be implemented with much greater specificity, especially with regard to classroom instruction. The objectives stated in Step Two can be particular skills such as asking questions, using certain responsive acts more frequently, learning how to modify patterns of instruction from one situation to the next, or learning how to write the objectives of instruction in denotative language. This means that the standards of performance which are set are likely to be unique to each project, even though they can be assessed with considerable objectivity.

Teachers, of course, will vary widely in their knowledge of new assessment procedures and their skill in using them. Evaluation projects designed by less experienced teachers will doubtless make use of simple assessment procedures. We can hope that this kind of knowledge and skill will improve with experience so that more extensive evaluation projects may occur as teachers' competence increases.

Nonteaching Certificated Personnel

The nonteaching certificated personnel who work most closely with teachers to improve both professional performance and instruction are likely to be supervisors and principals. We propose that it is their adjunctive duty to promote the evaluation of instruction in ways that will contribute to the improvement of education.

Research

Studies of successful and unsuccessful attempts to improve public school education by means of various innovative programs have indicated that the school administrator plays a key role in the change process. Brickell (1964, pp. 503-505), surveying new instructional programs in New York State, noted that introduction of change depended almost exclusively upon administrative initiative. The most successful innovations were those which provided the most elaborate support to teachers as they attempted to change their instructional practice. Fox and Lippitt (1964, p. 297) found a much higher probability of instructional improvement when the efforts of the classroom teacher were accompanied by informed and sympathetic support from both school administrators and professional colleagues. They further noted that channels of communication within a school were needed to enable innovative teachers to share the results of their efforts with colleagues who were facing similar classroom problems. If such channels existed, teachers learned from each other.

California's experience with Title III programs (Johnson, 1964, p. 171) provides additional evidence of the need for administrative leadership. Fifty percent of the elementary schools in California reported extreme difficulty with in-service programs because of the inability of administrative and supervisory staff to provide effective leadership in the substantive content of new programs. As a result, districts were unable to complete their in-service plans or to conduct adequate evaluation of programs.

The indication of these and other, similar, studies (Grieder, 1969, p. 1046) is that improvement of control and learning environments is not apt to result from the isolated efforts of individual classroom teachers to change their instructional practices. Support systems are essential.

Recommendations

Since the role of the administrator is such a critical one, evaluation of adjunctive duties could begin with the following questions:

1. Does the principal or supervisor organize space, time, incentives, and resources for the improvement of teaching in sufficient quantity and quality so that a potent change environment for teachers (Flanders, 1970, pp. 336-344) comes into existence?

2. Does the principal or supervisor collect a variety of evaluation procedures and the necessary resource materials so that teachers with different interests and abilities can obtain access to an evaluation procedure that matches their knowledge and skills?

3. Does the principal or supervisor update his own competence regularly? What is the most recent new skill he has acquired? When and how did he acquire it?

4. Does the principal or supervisor facilitate the communication of teachers with their colleagues about their own attempts to improve instructional procedures? Which teachers communicate with whom? How many teachers within the school are involved in this communication network?

5. What outside resources and/or consultant services does the principal or supervisor make available to teachers to provide input about the particular problems they are trying to

solve?

It is quite probable that the support system in every school in California needs to be improved if implementation of the Stull Act is to lead to instructional improvement. Effective support systems will not be established without effort. It will be some time before they are fully developed, if indeed they ever are. But preliminary steps can be taken immediately.

Providing a Model of Inquiry. Supervisors and principals who design and conduct inquiry projects in which their own skills for supporting teacher evaluation are the objects of inquiry are likely to accomplish several desirable outcomes with the same project. Not only will they learn more about the evaluation of instruction and how they can support it, but even more important, they will provide a model of self-assessment and self-development for the teachers with whom they work. The positive reinforcement effects of providing a model of behavior, which Bandura and his colleagues at Stanford have reported (Bandura and Kupers, 1963), are more likely to be realized if the steps of inquiry that supervisors and principals carry out are similar, if not identical, to the procedures teachers will use. Consider the following parallels:

Inquiry Projects*

	Principals or Supervisors	Teachers
1.	Identify evaluation activities you would like your teachers to carry out.	Identify patterns of pupil behavior you would like your pupils to exhibit.
2.	Identify at least two different patterns of super-	Identify at least two different patterns of teacher

---

*A sample plan for a teacher inquiry project is included in Appendix A.

	Principals or Supervisors	Teachers
	visory behavior that you think will support the desired teacher behavior.	behavior that you think will support the desired pupil behaviors.
3.	Practice these supervisory patterns in order to learn them and to develop observation schemes for analyzing them.	Practice these teaching patterns in microteaching and train a colleague to observe these patterns systematically.
4.	Try out both supervisory patterns in comparable circumstances and collect data to evaluate and compare the two patterns.	Teach with each pattern in comparable situations and collect observation data to show that the patterns were performed, together with other data to show pupil reactions.
5.	Analyze the data and decide which pattern of supervision works best.	Analyze the data and decide which teaching pattern is more effective.

The parallels in these outlines are meant to show how teachers, supervisors, and principals can evaluate their own professional performance by following almost identical steps of inquiry. To the extent that this is true, supervisors and principals can take the lead in demonstrating how to evaluate one's own performance under circumstances in which the teacher's behavior is not being evaluated. If the assessment of teachers includes the progress of their pupils, then why not base the evaluation of supervisors and principals partially on the quality and quantity of inquiry projects that are literally carried out by teachers?

Initiating Retraining Projects. Supervisors and principals can also initiate projects to develop new skills, and can illustrate the process of continuous learning for the teachers with whom they work. The administrator who is actively retraining himself while on the job demonstrates the acceptability, even desirability, of admitting the need to learn new skills.

The Kettering Foundation has developed a design for helping administrators retrain themselves. In the individualized Continuing Education (ICE) system, the school district supports the principal in an individually designed program of self-improvement that will contribute to desired improvements in the school and district (Kettering, 1970). An extension of the ICE program is called SPAR, Self-Performance Achievement Record. In this program the principal's self-development plans are laid out in a contract with specification of behavioral objectives, performance criteria, and acceptable evidence that these have been met (Olivero et al., 1972). Skill training for administrators and teachers could have the following parallels:

Competency Contracts*

Administrators	Teachers
1. Survey faculty communication channels devoted to improving instruction.	Record several segments of classroom interaction, analyze them, and identify the primary characteristics.
2. Determine changes in communication that have highest priority, and skills that will implement these changes.	Determine high priority teaching skills to be developed, depending on instructional goals.
3. Select a particular administrative or supervisory skill	Select a particular teaching skill to be developed;

---

*A sample competency contract is included in Appendix A.

	Administrators	Teachers
	to be developed; set a self-performance goal to be reached.	set a self-performance goal to be reached.
4.	Begin a program of training to develop the skill.	Begin a program of training to develop the skill.
5.	Collect post training data to determine whether self-performance goal has been reached.	Collect post training data to determine whether self-performance goal has been reached.

As in the case of inquiry projects, administrators and teachers can follow almost identical steps in planning, conducting, and evaluating a program of skill training. For an administrator who selects this route, one criterion of his effectiveness could be the number of teachers on his staff who have successfully completed some needed skill training through use of a competency contract.

Skill Training Materials. The principal who is ready to undertake a training program to develop new skills himself will find some training materials available. The table which follows, indicates some essential skills, training materials, and evaluation instruments. Additional descriptions of the training materials will be found in Appendix B.

Providing Outside Resources. A principal who is not prepared either to model inquiry processes or to demonstrate retraining processes for his staff may still fulfill his adjunctive duty to evaluate and improve instructional competency by providing support to his teachers in the form of outside resources. Outside consultants can be used to identify student entry behavior, desired student terminal behavior, and parameters of teacher behavior which are predictive of student success. Teachers can be observed by trained consultants or

	ADMINISTRATIVE AND SUPERVISORY SKILLS	TRAINING PACKAGES	ASSESSMENT MEASURES
1.	The supervisor will state behavioral objectives for his own performance as well as the performance of teachers and children.	Developing Effective Instruction (General Programmed Learning, Palo Alto) Behavioral Objectives (Southwest Cooperative Educational Laboratory, Albuquerque)	Evaluation instruments are contained in the training package.
2.	The supervisor will record data on teachers' classroom interaction, using an objective observational system.	Interaction Analysis (Far West Laboratory) Systematic and Objective Analysis of Instruction (Northwest Laboratory)	Evaluation instruments are contained in the training package.
3.	The supervisor will provide teachers with useful feedback and suggestions on the basis of his observations, in a supportive manner.	Systematic and Objective Analysis of Instruction (Northwest Laboratory)	Evaluation instruments are contained in the training package.
4.	The supervisor or administrator will be perceived by his staff as knowledgeable and supportive.	Service provides analysis of responses and suggestions for methods of improving skills.	Administrator Image Questionnaire (Educator Feedback Center, Western Michigan University)
5.	The administrator will demonstrate skill in using group decision making to determine instructional goals.	Determining Instructional Purposes (Far West Laboratory, Educational Management Program)	Self-evaluation instruments are included in the training package.

technicians and their goal-directed behavior can be reinforced.

Two examples of effective systems of this type are the Quality Assurance Specialist Program (QAS), developed by the Southwest Cooperative Educational Laboratory in Albuquerque, and the Spaulding Behavior Modification System (Spaulding, 1971). The QAS program is aimed at pupil achievement in learning English as a second language. The Spaulding program is aimed at increasing independent, self-directed behavior of pupils. Both of these systems have been effective in changing teachers' behavior. Both could serve as prototypes for California schools.

In this type of procedure the principal or supervisor is not providing the training or evaluation himself, but he is arranging to have outside consultants work with his teachers to achieve a specific goal of the school. The principal who utilizes this type of support system might be evaluated on the basis of the general improvement shown by the staff in the instructional skills being reinforced.

Certificated Teaching Personnel

The adjunctive duties of teachers pertaining to evaluation of professional performance are basically similar to those of principals and supervisors. We propose that it is the teacher's adjunctive duty to: (1) evaluate and improve his own performance; and (2) assist his colleagues in the evaluation and improvement of their performance.

The processes by which the teacher can evaluate and improve his own performance have already been discussed in illustrating the parallel procedures that could occur when a supervisor models an approach to inquiry or demonstrates an approach to retraining. The effectiveness of inquiry projects has been demonstrated in one study that involved the use of operant conditioning procedures and labeled the process "precision teaching self-evaluation" (Neale, 1971). Teachers identified their immediate and long-range goals, collected baseline data on their own behavior and on pupil behavior in microteaching situations, planned a strategy to bring about desired changes in pupil behavior, and tested the effectiveness of the strategy in further microteaching situations. They were able to change both their own behavior and the behavior of their pupils.

While the process of inquiry or of competency contracting may be the same for both teachers and principals, the skills that teachers need to learn may differ markedly. Self-instructional training materials available to develop new teaching skills will be discussed later in this paper, in the section on proper control and suitable learning environment.

The process of assisting one's colleagues to evaluate and improve their performance was mentioned in the explanation of inquiry for teachers; it merits some additional discussion here. The Fox and Lippitt study (1964, p. 296) cited earlier stressed the need for innovative teachers to share their problems and discoveries with colleagues. Schmuck reported on alternative intervention strategies designed to help teachers improve informal group processes in the classroom, and noted that an effective procedure was to have pairs and trios of teachers plan and conduct inquiry projects (Schmuck, 1971, pp. 33 ff.). The Teachers College, Columbia University, Preservice Program has made extensive use of peer feedback as student teachers are trained to use a number of different instructional models in classroom interaction with children. The peer feedback is effective in helping student teachers learn to display measurably different patterns of teaching behavior as they play different instructional roles (Joyce et al., 1972).

At least two school districts in California have experimented successfully with dyadic peer relationships for the support of self-improvement programs (Wallace, 1972; S. L. Fox, 1972). One advantage of a support system involving peer observation and evaluation is that the teacher is less apt to "put on an act" in order to achieve a high rating, and is more apt to learn from the evaluation process. Peer feedback may be less threatening than evaluation by one's employer. As an added bonus, the teacher who is assisting a colleague by observing and providing feedback is also learning himself as he develops new perceptions of the teaching act.

The assessment of a teacher's performance of adjunctive duties could include consideration of the quality and quantity of his own inquiry projects. It should also give some weight to the quality and quantity of his assistance to his colleagues in the conduct of their inquiry projects.

Summary on Adjunctive Duties

The adjunctive duties most closely related to assessment of proper control and learning environment are those which involve evaluation and improvement of one's own professional performance and assistance to one's colleagues in evaluating and improving their professional performance. Three possible organizational procedures for carrying out these adjunctive duties, and consequently for carrying out the assessment of professional performance, have been identified: (1) use of inquiry projects; (2) use of competency contracts; and (3) use of outside consultants. These three procedures can be used separately or in conjunction with each other. Each procedure has been tested in at least one pilot project.

It has been suggested that the role of nonteaching certificated personnel in providing proper control and suitable learning environment be evaluated in part in terms of the quality and quantity of inquiry projects carried out by teachers and the number of teachers successfully completing competency contracts. The assessment of a teacher's performance of adjunctive duties could be evaluated in part by the quality and

and quantity of his own inquiry projects and competency contracts and by the quality of his assistance to his colleagues in their attempts to improve professional performance.

Evaluation of Proper Control and Suitable Learning Environment

We have defined proper control as the teacher's ability to create and maintain a climate in which the intended learning can occur without undue interruption. A suitable learning environment is said to exist when students are interested in coming to class, look forward optimistically to the work involved, and obtain a sense of satisfaction from participating, especially in terms of self-respect and self-confidence. These definitions represent goals that probably cannot be achieved for all students all of the time. Yet working toward these goals as if they can be accomplished is a crucial element of professional teaching. It is our position that standards for assessment of teachers' achievement of these goals should refer to identifiable levels of performance associated with particular teaching skills and should be selected and defined by the teacher whose performance is to be evaluated.

Earlier in this paper, we identified four criteria to be met if assessment of proper control and learning environment, viewed in terms of improvement of instruction, is to be considered valid and feasible. These criteria are agreement between teacher and supervisor on instructional goals; some research evidence linking teacher behavior with control and learning environment; measurement instruments that can be utilized by local school personnel as well as outside consultants; skill-training materials, at least some of which are self-instructional.

Three sources of data for evaluating proper control and suitable learning environment are teacher planning activities, student attitudes and perceptions, and the verbal interaction of teachers and pupils. Each of these data sources will be considered in light of the above criteria, and appropriate organizational procedures for the assessment of each will be identified.

Planning

On the basis of common sense we expect that classroom control and the learning environment are better when a teacher plans for instruction than when he does not plan. Planning for adaptation of curriculum materials to fit the needs of a particular class and particular children within a class may be taken as evidence that a teacher is attempting to provide proper control and a suitable learning environment. But can it be taken as evidence that a teacher is achieving proper control and a suitable learning environment? We think not.

Criteria for Valid and Feasible Assessment

Let us consider teacher planning activities in terms of the four criteria restated above. The first step in assessment is to reach agreement between a teacher and supervisor on instructional goals. This agreement is likely to be facilitated by the statement of instructional goals in terms of specific pupil behavior (a topic dealt with in another paper at this conference). Thus, one criterion of effective planning is the teacher's ability to specify objectives in behavioral terms. When specific objectives have been agreed upon, teacher performance can be assessed in terms of how well other planning activities contribute to achievement of those objectives.

Research Evidence. We know of no research projects which have the purpose of showing that making plans for instruction (versus not making plans) results in the improvement of desirable learning outcomes. Perhaps such studies do not exist because the answer seems self-evident. Relevant research with regard to planning does provide some suggestions concerning the particular aspects of teacher planning that might be most closely related to control and learning environment. Berlyne (1960) reports that novelty, complexity, and "surprisingness" all increase the attention of students in the classroom. Other studies indicate that providing a variety of learning activities reduces "satiation" and is conducive to appropriate classroom behavior (Kounin, 1966). Biddle (1967) notes that classroom arrangements in which the teacher occupies center stage and pupils are encapsulated in small areas tend to be settings of some boredom for pupils. These studies suggest the possibility of evaluating the teacher's planning of both the physical arrangement of the classroom and the instructional materials available for use, with variety as an important criterion in both instances.

Planning for variety of instructional materials makes more sense when class formations and plans for the role of the teacher are also taken into consideration. For example, as work begins on a unit of study, the first task is to identify individual learning goals for students. There is a stage of teacher-student planning that calls for an appropriate class formation and particular patterns of teaching behavior. Later on, as work gets under way, another class formation, a shift in the role of the teacher, and the use of a variety of learning materials can be expected. Some suggestions about expected changes in teaching behavior as individual learning goals are clarified in mathematics and social studies classes have been reported by Flanders (1970, p. 327). Joyce and Weil (1972) identify and discuss different teaching models or strategies as appropriate for different instructional objectives. Taba (1966) also proposes particular changes in teaching behavior designed to support and encourage skill in inductive thinking.

We might note, in passing, that changes in instructional materials do not necessarily ensure complementary changes in patterns of teaching behavior. There is evidence that modern mathematics curricula (Wright, 1967) and the new physics curriculum designed by the Physical Science Study Committee (Moore, 1968) failed to promote more independent patterns of student participation primarily because old patterns of teaching behavior were continued when new patterns were required.

Measurement Instruments and Training Materials. There is no available supply of tested measurement instruments for the assessment of teacher performance in planning for variety of classroom arrangements and instructional materials. There are, however, both assessment instruments and training materials to deal with teacher performance in stating behavioral objectives. (These were noted in an earlier section.) We have been unable to locate any training materials dealing with skills in planning for variety of classroom arrangements and use of instructional materials, though a training format for such materials has been proposed (Morine, 1972). Some books explicating the process of planning for the use of different instructional models are available (Joyce and Weil, 1972; Morine and Morine, 1973), as are self-instructional materials to develop the interactive skills associated with such models (Joyce, Weil, and Wald, 1972).

Skill-training materials related to two other facets of teacher planning have been produced. The Far West Laboratory's "Minicourse on Independent Learning" deals with teacher-pupil planning of independent learning contracts and emphasizes teaching behaviors related to providing for individual differences in academic interests and self-direction, both of which are relevant to control and learning environment. The training package on team teaching developed by the Austin, Texas, Research and Development Center deals with teacher-teacher planning and emphasizes behaviors that may be most closely related to the adjunctive duty of providing support and assistance to one's colleagues.

Appropriate Assessment Procedures

Given this state of affairs, it seems evident that improvement of some planning skills is an achievable goal but that assessment in terms of normative standards would be indefensible. We recommend that assessment of planning proceed through use of competency contracts. A particular planning skill to be assessed can be identified jointly by teacher and administrator. Base-line data on teacher performance of this skill can be gathered by some objective means mutually agreed upon. Performance criteria that represent skill development in a particular direction can then be specified. Training materials would be made available to the teacher. Final assessment would be based upon attainment of the specified performance level.

Student Attitudes

The point was made earlier that suitable learning environment is essentially a child-centered concept, referring to arrangements which maximize a child's opportunity to exploit a situation for the purpose of learning. It follows that students' attitudes toward the learning environment will be an important indicator of the suitability of that environment for them. Student attitudes toward themselves and toward their peers are apt to be related to their patterns of social behavior, and thus may have an effect on classroom control.

Criteria for Valid and Feasible Assessment

Agreement on attitudinal goals that may be related to control and the learning environment can probably be achieved by most teachers and supervisors. Most educators support such goals as these: the student will value learning, enjoy school, believe in his own ability to achieve, accept responsibility for helping his peers, value and display independence, and be willing to accept group decisions and obey group rules. Where disagreements frequently occur is in the identification of instructional procedures to implement these goals.

Research Evidence. It is encouraging to note that a large group of studies, altogether more than 20,* showed that when teaching behavior was more responsive, students had more positive attitudes toward the teacher and toward learning, often learned more (as shown by achievement tests), and often showed more desirable patterns of behavior in the classroom.

There are also a few studies relating student attitudes toward themselves and their peers to classroom control and learning environment. Wide sociometric dispersion in a group is related to higher group cohesiveness and stronger group norms supporting educational goals (Schmuck, 1966). Low sociometric status is related to negative self-esteem and hostility toward others (Lippitt and Gold, 1959). Teacher praise is likely to increase the sociometric status of the particular students toward whom it is directed (Flanders and Havumaki, 1960). While the research in these areas is rather meager, it does suggest that sociometric patterns may indicate the existence of problems in the area of control and learning environment. It does not suggest many solutions to such problems or point strongly to a particular direction for teacher improvement.

Measurement Instruments and Training Materials. A variety of instruments is available to measure student perceptions of the teacher, their peers, and themselves. Twenty-three different diagnostic forms for obtaining student perceptions are discussed by R. S. Fox, Luszki, and Schmuck (1966) in a very interesting little pamphlet called Diagnosing Classroom Learning Environments. The development of microteaching at Stanford University included a student reaction form called the "Stanford Teacher Competence Appraisal Guide" (STCAG) designed to obtain student perceptions of teaching performance. The "Michigan Student Questionnaire" has been used in several studies relating pupil attitudes to teaching behavior, and has demonstrated reliability (Flanders et al., 1969, pp. 169-179). The Educator Feedback Center of Western Michigan University provides a service for teachers that includes distribution and analysis of questionnaires dealing with pupil perception of teacher effectiveness and preparation of a teacher-image profile (see Appendix B for additional information). An instrument designed by Gage collects pupil descriptions of their actual and their ideal teachers and permits the teacher to compare himself with a "pupil-designed" model teacher (Gage et al., 1960).

---

*Flanders, 1965 (four separate studies); Flanders, 1969 (three separate studies); Filson, 1957; Amidon and Flanders, 1961; Schantz, 1963; LaShier, 1965; Johns, 1966; Morrison, 1966; Snider, 1965; Soar, 1966; Birkin, 1967; Emmer, 1967; Furst, 1967; Measel, 1967; Pankratz, 1967; Powell, 1968; Samph, 1968; and Weber, 1968.

The reactions of students who cannot respond by using a paper-and-pencil form can be obtained by asking them or by using an observation system that includes nonverbal student observation. Fairly reliable ratings for younger age levels have been reported by Soar (1966) on such matters as "at work/not at work," "interest/boredom," "participating/not participating."

While assessment measures for student attitudes exist, there are no training materials specifically designed to help teachers improve pupil attitudes toward school, toward each other, or toward themselves. Thus assessment of student attitudes will not lead easily into a skill-training program.

Appropriate Assessment Procedures

Assessment of student attitudes can lead to use of inquiry projects. When data on student attitudes have been collected, the teacher can identify alternative patterns of teaching behavior that might support a change in pupil attitudes, practice these teaching patterns with different groups of pupils, and collect further data from students to determine which pattern was more effective in bringing about the desired change. Final assessment of teaching performance would be based upon the teacher's ability to plan, conduct, and evaluate the inquiry project, as well as the ability to learn new patterns of behavior and to bring about some change in pupil behavior. Possible designs for inquiry projects are discussed in the next section.

Interactive Behavior

The interactive events that occur between a teacher and his students, as well as those that occur among students, create a major part of the learning environment and also reveal patterns of classroom control. In fact, as students "interact" with instructional materials, which are inanimate objects, the adaptation of these materials for purposes of learning--the "humanizing" of these materials, so to speak--is created by what teachers and students say to each other and how they coordinate their actions. It follows that an analysis of interactive behavior should give us considerable insight into the qualities of the learning environment and the processes of control. This kind of evaluation can be based on many different systems of interaction analysis. Perhaps the first analysis of this kind was made in 1912 by Stevens (1970), who studied teacher questions from written transcripts recorded in stenographic shorthand. An extensive inventory of the observation procedures now available can be found in Simon and Boyer's Mirrors for Behavior (1970), in which more than 80 systems for encoding interactive events are reported.

One general conclusion from all of the research which uses some form of interaction analysis is that, when experienced teachers or college students take the time to analyze their own patterns of verbal interaction, they are likely to change these patterns. Many different changes have been reported, but among those which occur most frequently are less teacher talk, more responsive teacher talk (e.g., reacting to ideas students express), and less teacher initiation (e.g., lecturing, giving directions, and criticizing students). This generalization is supported, to one degree or another, by no less than 13 different research projects.* One possible reason for these consistent results is that interaction analysis, with proper feedback, helps a teacher compare what actually happened during classroom interaction with what he wanted to have happen. When a teacher analyzes his own chain of interactive events, he is more likely to discover appropriate opportunities to react to students, and this increases the average incidence of responsive acts. Additional explanation has been proposed by Wagner (1971), whose study suggests that practice in discriminating among responsive acts may be more important than practice in performing these acts.

As noted earlier, there is also a large group of studies linking responsive teaching behavior to positive student attitudes, higher pupil achievement, and more desirable patterns of classroom behavior. Of course, reservations and exceptions have been made to these two rather sweeping generalizations. For example, it has been remarked that not all teachers can benefit equally from the study of their own interactions (Zahn, 1965, p. 297), cause and effect are not yet clearly established (Flanders, 1970, p. 426), the relationships may actually be curvilinear, which means a teacher can become too responsive

---

*At the preservice level, Lohman, Ober, and Hough, 1967; Moskowitz, 1967; Kirk, 1963; Furst, 1967a; Finske, 1967; Bondi, 1969. At the inservice level, Flanders, 1963; Emmer, 1967; Hill, 1966; Jeffs, 1968; and Soar, 1966. Others have reported similar findings since this survey was made, for example Borg et al. (1970).

(Soar, 1966), and not all classes, subject matter, age levels, and ways to quantify variables are properly represented in the research reported (Rosenshine and Furst, 1971, pp. 37 ff.). However, the line of research is still impressive and deserves attention in terms of implementing the Stull Act.

Criteria and Inquiry Design

The problem of evaluation is to obtain information based on fairly objective criteria and then to use this information to make intelligent comparisons that, in turn, lead to valid conclusions. In research we say that variables should be carefully quantified and the data should be analyzed within a logical research design. Evaluation under the terms of the Stull Act may only approach the quality of carefully conceived research, yet the need for attention to the principles of good research is inescapable.

Criteria for Interactive Events. The criteria of evaluation when some system of interaction analysis is used will depend on the nature of the category system. In the Flanders ten-category system (1970, p. 197), the criteria can include ratios of different verbal statements made by the teacher (e.g., the i/i + d ratio); direct percentages of teacher talk, student talk, or silence/confusion; or subdivisions of the main categories. Other systems such as Parson's (1970) "Guided Self-Analysis" and Morine's (Morine, Spaulding, and Greenberg, 1971) discriminate different kinds of questions and different ways that a teacher can respond to student statements. The Coping Analysis Schedule for Educational Settings (CASES) and the Spaulding Teacher Activity Rating Schedule (STARS) by Spaulding (1971) discriminate types of pupil behavior and teacher use of positive or negative reinforcement.

Coding and analysis of total classroom interaction require extensive time and training. It is possible for evaluation to focus on one or two aspects of interaction and for coding to be limited to particular types of behavior. It is our recommendation that initial assessment of verbal interaction skills focus on teacher questions and directions, which occur just before pupils participate, and teacher responsive behaviors, which occur just after pupils participate (Flanders, 1972). Such assessment will require the following procedures:

1. A trained observer will encode interaction analysis data by visiting live classroom settings or will have access to an intelligible voice and/or video magnetic recording.

2. The encoded data will be tabulated or placed in a display which permits questions, directions, responsive statements, and criticism to be identified, along with various kinds of student statements. Additional ratings of the situation and data from student inventories can also be utilized.

3. Fairly simple ratios or summary percentages can be calculated in order to provide scores for subsequent comparisons.

The trick, if we can call it that, is to avoid superficial, unreliable indices which have questionable value and to focus instead on features of the interactive events which are more likely to have pervasive consequences for the learning environment. For example, counting particular kinds of words (such as an "I/Thou" ratio) may seem to be related to Martin Buber's observations and philosophy about personal relationships, but we do not have convincing evidence that such word counts are related to educational outcomes, either desirable or undesirable. On the other hand, the proportion of responsive teacher statements to all teacher statements (the i/i + d ratio) has been shown to be associated with positive student attitudes.

Inquiry Designs. Considerable attention should be given to the situations in which interaction analysis data are collected. The main problem is to choose an inquiry design which permits logical comparisons that are consistent with the purpose of the evaluation.

Early in this paper, one-time evaluations for the purpose of rating the quality of teaching were rejected. One reason for this recommendation is that a single visit is not likely to be representative of what normally occurs in a classroom. To secure more representative data may require at least six one-hour visits, well spaced (Flanders, 1970, p. 100) in order to collect reliable data. Samph (1968) has shown that the presence of an observer does influence interaction. The use of "standardized" lessons shows some promise of producing more reliable data (Finske, 1967), but more research will be necessary to demonstrate that such lessons represent average interaction. Most of us who conduct research using interaction analysis have discovered that it is not difficult for a teacher to put on an act which cannot be detected by most interaction analysis procedures. The final blow to one-time evaluation is that we simply do not have enough normative data to evaluate a teacher's performance in comparison with reasonably similar populations. Comparisons over time for the same teacher and the same class may be the best type of design for the diagnostic evaluation of the learning environment and the control processes.

One design could involve comparing interaction data with a model. A profile of interaction analysis data can be used to represent a particular model of teaching, such as a model of inquiry teaching. Examples of such a model can be found in Morine et al. (1971). A teacher can practice using this model and, on successive occasions, can compare his profile based on a specimen of interaction with the model display itself. Thus he can determine when his teaching matches the model. The primary purpose of this design is the improvement of instruction, rather than some kind of rating.

A second design would require a colleague or older student who is a trained observer to visit the class and obtain a general profile of interactive events at a first point in time. The teacher who was observed may identify some feature of this profile that he would like to change. He can then practice making these changes in microteaching or short lesson sessions. He would finally invite his colleague-observer back for a second observation. A comparison of the data collected at two different points in time can often lead to useful insights. Again, this design is directed primarily toward the improvement of instruction.

The five-step inquiry model for teachers outlined earlier in this paper is perhaps the most advanced level of inquiry in which one's own behavior is the object of study. It is more advanced because, in Step Two, the teacher proposes his own model. These five steps are discussed in considerable detail, including proposed systems for displaying the interaction data, in Analyzing Teaching Behavior (Flanders, 1970, Chapter 9).

Combining Interaction Analysis with Other Data-Collection Procedures

The generalization that teachers tend to change interaction patterns when they study and analyze their own verbal statements is most easily explained by considering "feedback" and the effects of feedback on a behavior system. More adequate feedback to teachers is likely to occur if more adequate information is obtained. One way in which this can be accomplished is to help teachers develop the skills necessary to collect other kinds of information as well as information about interaction.

Two kinds of data, interaction data and student perception data, are essential to adapting teaching methods to one's own class. The interaction data permit a teacher to determine the extent to which his interaction matched his intentions. The data on student perceptions help to associate student reactions with the patterns of instruction. The combination permits inferences about how well a given strategy works for a particular class.

Regular Feedback. Incidental feedback to the teacher goes on constantly. Systematic feedback occurs less frequently. Change is more likely to result from the latter because it is planned.

An effective, easy way to obtain feedback from students as part of an inquiry project is to ask them for their opinions. It may be both more tactful and more effective to divide the class into groups of four or five students, ask them to form opinions about some phase of instruction, and then have a "summarizer" report for the subgroup, not naming individual students. After opinions are rewritten to reflect the various subgroup discussions, a show of hands to indicate the amount of agreement and the strength of conviction helps to quantify class opinion.

Short forms for written or scaled student reaction which take only a few minutes to fill out can be created by most teachers. Many examples are shown by Fox, Luszki, and Schmuck (1966).

Feedback on interactive behavior is not quite so readily obtained. Time, space, incentive, and resources are necessary to learn to encode and decode classroom interaction. This kind of activity is not easily accomplished if it is an extra burden added to the load of an already busy teacher or supervisor.

One deterrent to learning how to use interaction analysis systems may be exaggerated demands for reliability of observations. The authors of this paper have worked with teachers as they learn some form of interaction analysis. We have found that the standards of reliability required for research projects need not necessarily be reached in order to benefit instruction. When two teachers disagree on how to code an event and they discuss their differences, it probably helps them to improve their discrimination.

Other factors will affect the success of training attempts. At the Far West Laboratory we have some unpublished information from our field tests to indicate that teachers who work alone are less likely to complete training in interaction analysis than are teachers who work with one or more partners. The Northwest Regional Laboratory's materials (see Item D, p. 217, Appendix B) are purposely designed to fit the format of a weekend workshop for a reasonably sized group of teachers. If such a workshop is not feasible, self-instructional training packages should probably be used by several teachers in the same school for maximum effectiveness. There are quite a few training materials for learning some form of interaction analysis. Descriptions of these can be found in Appendix B.

Skill Training

Any assessment of teacher performance that is aimed at improvement of instruction will eventually lead to the identification of new teaching skills that should be acquired. Not all teachers will be ready to conduct inquiry projects or write competency contracts. For many the route to improvement of instructional skills will be the use of training packages that identify skills to be learned, provide carefully designed materials to assist the teacher in skill acquisition, and include assessment measures to determine the extent of learning that has occurred.

It is probable that many teachers will need training in stating behavioral objectives before they can write competency contracts. Training in coding interactive behaviors will be essential for many teachers before they can begin inquiry projects. Some teachers will move rapidly from skill-training projects to more complex assessment procedures. Others will remain at the level of skill training for longer periods of time.

One of the more powerful approaches to the improvement of interactive teaching skills is the development of "minicourses."

These are products of the Far West Laboratory for Research and Development (Borg, W. R. et al., 1970). Minicourses combine the techniques of microteaching with Bandura and Kupers' (1963) work on providing models of behavior. A minicourse provides a coordinator's manual, a teacher's manual, and color films or video recordings to demonstrate particular teaching skills, all organized around a series of lessons. Individualized training is possible for a single teacher or a group of teachers. Each minicourse is the product of an extensive sequence of product development and evaluation. Evidence of product effectiveness can be secured from the Far West Laboratory.

Other regional laboratories and university centers are currently developing and testing instructional materials designed to improve particular teaching skills. Reference to these materials can be found in this chapter and also in Appendix B. The advantage of most of these products is that they are designed to assist those teachers who do not care to design their own self-development programs. The procedures of interaction analysis, discussed in the previous section, require considerable teacher initiative and often call for the ability to invent a model of teaching (e.g., "What pattern of teaching behavior will support the desired pupil behavior?"). The self-contained minicourses and the instructional materials from laboratories and centers have the advantages that skills are usually specified in advance and that the materials specify a series of steps that will permit a trainee to improve his skill providing he follows the sequences of instruction.

Space does not permit an exhaustive list of the possible skill training tools that could contribute to the improvement of classroom control and learning environment. We include below a sampling of skills and training materials to indicate something of the range of possibilities that exists. (More detailed information is included in Appendix B.)

Teaching Skills	Training Packages
1. The teacher will state performance objectives for the instruction of his pupils.	Developing Effective Instruction (General Programmed Learning) Behavioral Objectives (Southwestern Cooperative Educational Laboratory)
2. The teacher will record data on his own classroom interaction using an objective observational system.	Interaction Analysis (Far West Laboratory) Interaction Analysis (Northwest Regional Laboratory)
3. The teacher will interact with pupils in ways that encourage high-level thinking skills.	Facilitating Inquiry in the Classroom (Northwest Regional Laboratory) Effective Questioning: Elementary Level (Minicourse 1, Far West Laboratory) Higher Cognitive Questioning (Minicourse 9, Far West Laboratory)
4. The teacher will organize the classroom for individualized instruction.	Organizing for Independent Learning (Minicourse 8, Far West Laboratory) Individualizing Instruction in Mathematics (Minicourse 5, Far West Laboratory) Teaching in IPI Mathematics (Research for Better Schools)
5. The teacher will increase the students' interest in learning.	Teaching Achievement Motivation (Education Ventures)
6. The teacher will expand the language and thought of young children.	Developing Childrens' Oral Language (Minicourse 2, Far West Laboratory)

Summary on Proper Control and Suitable Learning Environment

Assessment of proper control and suitable learning environment can proceed through collection of data related to teacher planning, student attitudes, and interactive behavior of teachers and pupils. The use of such data for improvement of instruction can provide valid and feasible assessment procedures. Directed skill training, competency contracts, and inquiry projects are all appropriate means of assessment and improvement at this stage in our development of techniques for evaluation of teacher effectiveness. Inquiry projects can involve comparisons of behavior with a model, comparisons of behavior with stated behavioral objectives, and comparisons of behavior over time. Peer cooperation in collection of interactive data is apt to contribute to development of skills in coding interactive behavior and therefore to lead to greater behavior change.

A Beginning

The educational community of California has been issued an ultimatum to evaluate the performance of certificated personnel throughout the state. We do not have the necessary knowledge base or the required measurement techniques and tools to respond to this ultimatum by instituting a system of rigid evaluative criteria for purposes of retention or merit pay. We do have the opportunity to respond by instituting systems of assessment that can contribute to the improvement of instruction.

Improvement of the educational process is a goal to which we all aspire. Perhaps we wish that we had not been placed under such immediate pressures to achieve this goal as the Stull Act seems to exert. But if we marshal our resources, we can make an important initial attack on the problems of accountability, effective teaching, and improvement of instruction.

There are those who will view the Stull Act as an obstruction or a dead end. We can make it a beginning.

References

Amidon, E. J. and Flanders, N. A. "The Effects of Direct and Indirect Teacher Influence on Dependent-Prone Students Learning Geometry." Journal of Educational Psychology, 1961, 52, 286-291.

Bandura, A. and Kupers, C. J. "Transmission of Patterns of Self-Reinforcement through Modeling." Journal of Abnormal and Social Psychology, 1963, 69, 1-9.

Berlyne, D. E. Conflict, Arousal, and Curiosity. New York: McGraw-Hill, 1960.

Biddle, B. J. and Adams, R. S. An Analysis of Classroom Activities. Columbia: University of Missouri, 1967.

Birkin, T. A. "Toward a Model of Instruction Processes." Unpublished manuscript, School of Education, University of Birmingham, England, 1967.

Bondi, J. C., Jr. "The Effects of Interaction Analysis Feedback on the Verbal Behavior of Student Teachers." Educational Leadership, 1969, 26, 794-799.

Borg, W. et al. The Minicourse: A Microteaching Approach to Teacher Education. 8701 Wilshire Boulevard, Beverly Hills, CA: Macmillan Educational Services, 1970.

Brickell, H. M. "State Organization for Educational Change: A Case Study and a Proprosal." In M. B. Miles (Ed.), Innovation in Education. New York: Teachers College, Columbia University, Bureau of Publications, 1964.

Emmer, E. T. "The Effect of Teacher Use and Acceptance of Student Ideas on Student Verbal Initiation." Unpublished doctoral dissertation, University of Michigan, 1967.

Filson, T. N. "Factors Influencing the Level of Dependence in the Classroom." Unpublished doctoral dissertation, University of Minnesota, 1957.

Finske, Sister M. J. "The Effect of Feedback through Interaction Analysis on the Development of Flexibility in Student Teachers." Unpublished doctoral dissertation, University of Michigan, 1967.

Flanders, N. A. *Helping Teachers Change Their Behavior*. (Final Report, N.D.E.A. Projects 1721012 and 7-32-0560-171.0, U. S. Office of Education). School of Education, University of Michigan, 1963. (Out of print)

Flanders, N. A. *Teacher Influence, Pupil Attitudes, and Achievement*. Cooperative Research Monograph No. 12 (OE-25040), U. S. Office of Education. School of Education, University of Michigan, 1965. (Out of print)

Flanders, N. A. *Analyzing Teacher Behavior*. Reading, Mass.: Addison-Wesley, 1970.

Flanders, N. A. "Basic Teaching Skills Derived from a Model of Speaking and Listening." Paper presented at Basic Skills Seminar, American Educational Research Association Annual Meeting, Chicago, 1972. (Available from author.)

Flanders, N. A. and Havumaki, S. "The Effect of Teacher-Pupil Contacts Involving Praise on the Sociometric Choices of Students." *Journal of Education Psychology*, 1960, *51*, 65-8.

Flanders, N. A. et al. "Teacher Influence Patterns and Pupil Achievement in the Second, Fourth, and Sixth Grade Levels." Cooperative Research Project No. 5-1055 (OE-4-10-243), U. S. Office of Education. School of Education, University of Michigan, December 1969. (Available from Far West Laboratory, San Francisco, California.)

Fox, R. S. and Lippitt, R. "The Innovation of Classroom Mental Health Practices." In M. B. Miles (Ed.), *Innovation in Education*. New York: Teachers College, Columbia University, Bureau of Publications, 1964.

Fox, R. S., Luszki, M. B., and Schmuck, R. *Diagnosing Classroom Learning Environment*. Chicago: Science Research Associates, 1966.

Fox, S. L. "The sponsor/colleague system." *California School Boards*, 1972, *31*, 19.

Furst, N. "The Effects of Training in Interaction Analysis on the Behavior of Student Teachers in Secondary Schools." In E. J. Amidon and J. B. Hough (Eds.), *Interaction*

Analysis: Theory, Research, and Application. Reading, Mass.: Addison-Wesley, 1967a.

Furst, N. "The Multiple Languages of the Classroom: A Further Analysis and a Synthesis of Meanings Communicated in High School Teaching." Unpublished doctoral dissertation, Temple University, 1967b.

Gage, N. L. et al. "Equilibrium Theory and Behavior Change: An Experiment in Feedback from Pupils to Teachers." Urbana, Ill.: Bureau of Educational Research, University of Illinois, 1960.

Grieder, C. "Public School Administration." In R. L. Ebel (Ed.), *Encyclopedia of Educational Research*. American Educational Research Association, 1969.

Hill, W. M. "The Effects on Verbal Teaching Behavior of Learning Interaction Analysis as an In-Service Education Activity." Unpublished doctoral dissertation, Ohio State University, 1966.

Jeffs, G. A. et al. "The Effects of Training in Interaction Analysis on the Verbal Behavior of Teachers." Mimeographed research report. Ed. W. Clark High School, Las Vegas, Nevada, 1968.

Johns, J. P. "The Relationship between Teacher Behavior and the Incidence of Thought-Provoking Questions by Students in Secondary Schools." Unpublished doctoral dissertation, University of Michigan, 1966.

Johnson, D. W. "Title III and the Dynamics of Educational Change in California Schools." In M. B. Miles (Ed.), *Innovation in Education*. New York: Teachers College, Columbia University Bureau of Publications, 1964.

Joyce, B. R. and Weil, M. *Models of Teaching*. Englewood Cliffs, N. J.: Prentice-Hall, 1972.

Joyce, B. R., Weil, M. and Wald, R. *Three Teaching Strategies for the Social Studies*. Chicago: Science Research Associates, 1972.

Joyce, B. R. et al. "Models of Teaching as a Paradigm for Teacher Education." Paper presented at the American Educational Research Association meeting, Chicago, 1972.

Kettering Program Associates. "Individualized Continuing Education for School Administrators--One Approach." A CFK Ltd. Occasional Paper. Englewood, Colorado, 1970.

Kirk, J. "Effects of Teaching the Minnesota System of Interaction Analysis to Intermediate Grade Student Teachers." Unpublished doctoral dissertation, Temple University, 1963.

Kounin, J. S. et al. "Managing Emotionally Disturbed Children in Regular Classrooms." Journal of Educational Psychology, 1966, 57, 1-13.

LaShier, W. S. "An Analysis of Certain Aspects of the Verbal Behavior of Student Teachers of Eighth Grade Students Participating in a BSCS Laboratory Block." Unpublished doctoral dissertation, University of Texas, 1965.

Lippitt, R. and Gold, M. "Classroom Social Structure as a Mental Health Problem." Journal of Social Issues, 1959, 15, 40-58.

Lohman, E., Ober, R. and Hough, J. B. "A Study of the Effect of Pre-Service Training in Interaction Analysis on the Verbal Behavior of Student Teachers." In E. J. Amidon and J. B. Hough (Eds.), Interaction Analysis: Theory, Research, and Application. Reading, Mass.: Addison-Wesley, 1967.

McDonald, S. C. "The Passage of AB293, The Stull Act of 1971." Unpublished report. (Available from Far West Laboratory, Berkeley, California.)

Measel, W. W. "The Relationship between Teacher Influence and Levels of Thinking of Second Grade Teachers and Pupils." Unpublished doctoral dissertation, University of Michigan, 1967.

Moore, J. R. "An Analysis of Teacher and Pupil Verbal Behavior and Teacher Procedural and Evaluative Behavior in Relation to Objectives Unique to the PSSC and Non-PSSC Physics Curricula." Unpublished doctoral dissertation, University of Michigan, 1968.

Morine, G. "Basic Teaching Skills: Identification of Preactive Skills." Paper presented at Seminar on Basic Skills, American Educational Research Association meeting, Chicago, 1972.

Morine, G., Spaulding, R.and Greenberg, S. Discovering New Dimensions in the Teaching Process. Scranton, Pa.: Intext Educational Publishers, 1971.

Morine, H. and Morine, G. Discovery: A Challenge to Teachers. Englewood Cliffs, N.J.: Prentice-Hall, 1973.

Morrison, B. M. "The Reactions of Internal and External Children to Patterns of Teaching Behavior." Unpublished doctoral dissertation, University of Michigan, 1966.

Moskowitz, G. "The Attitudes and Teaching Patterns of Cooperative Teachers and Student Teachers Trained in Interaction Analysis." In E. J. Amidon and J. B. Hough (Eds.), Interaction Analysis: Theory, Research, and Application. Reading, Mass.: Addison-Wesley, 1967.

Neale, D. C. Precision Teacher Self-Evaluation. University of Delaware, Cooperative Center for Educational Development and Services, 1971.

Olivero, J. L. et al. "The Self-Performance Achievement Record (SPAR)." A CFK Ltd. Occasional Paper. Englewood, Colorado, 1972.

Pankratz, R. "Verbal Interaction Patterns in the Classrooms of Selected Physics Teachers." In E. J. Amidon and J. B. Hough (Eds.), Interaction Analysis: Theory, Research, and Application. Reading, Mass.: Addison-Wesley, 1967.

Parsons, T. W. Guided Self-Analysis System for Professional Development: Guidelines for Implementation. G.S.A. Education Series. Berkeley, 1970.

Powell, E. R. "Teacher Behavior and Pupil Achievement." Paper presented at American Educational Research Association meeting, Chicago, 1968.

Rosenshine, B. and Furst, N. "Research on Teacher Performance Criteria." In B. O. Smith (Ed.), Research in Teacher Education: A Symposium." Englewood Cliffs, N.J.: Prentice-Hall, 1971.

Samph, T. "Observer Effects on Teacher Behavior." Unpublished doctoral dissertation, University of Michigan, 1968.

Schantz, B. M. B. "An Experimental Study Comparing the Effects of Verbal Recall by Children in Direct and Indirect Teaching Methods as a Rool of Measurement." Unpublished doctoral dissertation, Pennsylvania State University, 1963.

Schmuck, R. A. "Improving Classroom Group Processes." In R. A. Schmuck and M. B. Miles (Eds.), Organization Development in Schools. Palo Alto: National Press Books, 1971.

Schmuck, R. A. "Some Aspects of Classroom Social Climate." Psychology in the Schools, 1966, 3, 59-65.

Simon, A. and Boyer, E. G. Mirrors for Behavior. Philadelphia: Research for Better Schools, 1970.

Snider, R. M. A Project to Study the Nature of Effective Physics Teaching. Cooperative Research Report No. S-280, U. S. Office of Education. Cornell University, 1965.

Soar, R. S. An Integrative Approach to Classroom Learning. Final Report, Public Health Service Grant No. 5-R11 MH 01096 to the University of South Carolina and Grant No 7-R11 MH 02045 to Temple University. Temple University, 1966.

Spaulding, R. L. Educational Intervention in Early Childhood. Final Report, Vol. I. Durham Education Improvement Program. Durham, N.C.: Duke University, 1971.

Stevens, R. "The Question as a Measure of Efficiency in Instruction." Reprinted in Mirrors for Behavior, Vol. XIV, Simon, A. and Boyer, E. G. (Eds.). Research for Better Schools, Philadelphia, 1970.

Taba, H. Teaching Strategies and Cognitive Functioning in Elementary School Children. Cooperative Research Project #2404. San Francisco State College, February 1966.

Wagner, A. C. "Changing Teaching Behavior: A Comparison of Microteaching and Discrimination Practice." Unpublished doctoral dissertation, University of Michigan, 1971.

Wallace, C. E. "Self-Appraisal: Systems for Improvement." California School Boards, 1972, 31:14.

Weber, W. A. "Relationships between Teacher Behavior and Pupil Creativity in the Elementary School." Unpublished doctoral dissertation, Temple University, 1968.

Wright, M. E. "Teacher-Pupil Interaction in the Mathematics Classroom." Technical Report No. 67-5, Minnesota National Laboratory. Minnesota State Department of Education, 1967.

Zahn, R. "The Use of Interaction Analysis in Supervising Student Teachers." Unpublished doctoral dissertation, Temple University, 1965.

Appendix A

SAMPLE INQUIRY PROJECT PLAN AND COMPETENCY CONTRACT

Sample Plan for an Inquiry Project

Teacher's Name: Mary T.

School: Samuel P. Zellerbach Elementary School

The desired patterns of pupil behavior are:

pupils will listen to each other's ideas

The acceptable evidence that such patterns have been achieved will be:

pupils independently summarize, explain, or ask for clarification of the ideas expressed by others in the group

Two different patterns of teaching behavior that might support the desired pupil behaviors are:

1. I will "model" the desired pupil behavior by summarizing, expanding on pupil ideas, or asking for clarification, with the expectation that pupils will begin to imitate this behavior.
2. I will ask pupils to summarize, explain, or ask questions about each other's ideas, with the expectation that after some practice they will respond in these ways without being asked.

The colleague who will work with the teacher on this project is: Jane M.

The method of observation will be: (Circle one)

"live" classroom visit listening to audio tapes

watching video tapes

The colleague will observe at least two practice lessons and will collect the following information:

Pattern I. a tally of the number of times I:

A. summarize a pupil's idea
B. expand on a pupil's idea
C. ask for clarification of a pupil's idea

Pattern II. a tally of the number of times I:

A. ask a pupil to summarize another pupil's idea
B. expand on a pupil's idea
C. ask pupils if they have any questions about a pupil's idea

The teacher will teach eight classroom lessons to two different groups using these patterns, and the colleague will observe four (2+2) classroom lessons and collect the following information:

same as above, plus a tally of the times that pupils independently:

A. summarize another pupil's idea
B. explain another pupil's idea
C. ask for clarification of another pupil's idea

The more effective teaching pattern will be identified by: deciding which pattern of teaching behavior results in more of the desired pupil behavior in the last set of classroom lessons observed.

Sample Competency Contract

Name: Mr. X
School: ABC Secondary School

Analysis of segments of classroom interaction shows:

A predominance of narrow questions, teacher direction, and teacher lecturing; very little pupil initiation

Related instructional goals are:

Pupils will develop skills in independent thinking

High priority skills to be developed are:

Asking higher order questions

Encouraging pupil planning and self-direction

Particular skill to be worked on now is:

Asking higher order questions

Available training materials are:

Book - Classroom Questions: What Kinds? by Norris Sanders

Workshop - Development of Higher Level Thinking Abilities, Northwest Regional Laboratory

Minicourse - Higher Cognitive Questioning, Far West Laboratory

Training material to be used:

Minicourse 9, Higher Cognitive Questioning

Performance goal to be reached:

Given training in Minicourse 9, Mr. X will increase use of higher order questions, so that analysis of a post-training segment of classroom interaction will show that over 50% of the questions asked are higher order questions, and that pupil answers are generally longer than teacher questions.

Expected time length:

Six weeks.

(Principal's Signature)

(Teacher's Signature)

Appendix B

AVAILABLE TRAINING MATERIALS*

1. Stating Behavioral Objectives

   A. Title: Behavioral Objectives Package

   Author or Developer: Southwestern Cooperative Educational Laboratory

   Source: SWCEL
   117 Richmond Drive NE
   Albuquerque, New Mexico 87106

   Cost: $132

   Materials: Slide-tapes; manual

   Description:

   The Behavioral Objectives Package contains slide-tape presentations and a workbook. Six lessons outline goals and objectives; the instructional program; the three domains (cognitive, affective, and psychomotor); entering behaviors; components of a behavioral objective; and how to write behavioral objectives. A pre- and posttest are included. Training in use of the package will be provided by the Southwest Cooperative Educational Laboratory.

   B. Title: Developing Effective Instruction

   Author or Developer: General Programmed Teaching

   Source: General Programmed Teaching
   P.O. Box 402, 424 University Avenue
   Palo Alto, California 94302

Cost:	
Complete workshop (15 units) approx.	$650.00
Student workbooks (each)	5.95

   (first five units may be purchased separately at $55.00 per unit)

   Materials: Audio-tape; filmstrip; workbook; monitor's manual

---

*Most of the information here is reprinted from B. Joyce, M. Weil, G. Morine, and R. Wald, Materials for Modules: A Classification of Competency-Oriented Tools for Teacher Evaluation (New York: Teachers College, Columbia University, 1971).

Description:

Entirely self-instructional course designed to teach principles and procedures for developing performance-based learning and instruction. Fifteen units deal with the following areas: basic principles of performance-based learning; general goals, affective objectives and cognitive objectives; main components of an objective; classification of objectives (verbal, discrimination, motor performance); criterion tests; determining entry level; entry level; tests; stimulus and response; one-way and two-way stimulus-response pairs; single and multiple discriminations; content analysis; stimulus-response pairs in chain activities; developing objectives and deriving content; programmed lesson plans and instructional media; validation.

The units assist the teacher in developing performance-based instruction on several levels: planning units and lessons; carrying out student contracts; creating programmed materials.

C. Title: Instructional Design: A Self-Directed Learning Program

Author or Developer: David P. Butts

Source: Research and Development Center for Teacher Education
University of Texas
Austin, Texas

Cost: Consult Source

Materials: Manual

Description:

The self-directed learning guide is designed as training in the skills for designing and planning instruction. It consists of 22 tasks organized around a single design or planning skill. For each task there is a pretest, a task activity, and a posttest. The task activity items require both multiple-choice and open-ended written responses. Most of the task activities are short and the entire sequence could be completed in about three to four hours. The task areas include behavioral objectives, task analysis, appraisal and assessment. matching instruction and materials to objectives, managing space and time, constructing instructional plans and initiating instruction.

D. Title: Vimcet Filmstrip - Tape Programs

Author or Developer: Vimcet Associates

Source: Vimcet Associates
P.O. Box 24714
Los Angeles, California 90024

Cost: $15.00 per program, filmstrip, and tape.
$ 2.00 per guide

Materials: Eighteen audio-tapes; filmstrips; instructor's manuals; utilization guide; related materials.
Supplementary texts: James Popham and Eva Baker, Establishing Instructional Goals: Planning an Instructional Sequence; Systematic Instruction. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Description:

This series consists of 18 filmstrip-tape programs concerned primarily, but not exclusively, with instructional design and planning skills. The titles of the programs are: Educational Objectives; Systematic Instructional Decision-Making; Selecting Appropriate Educational Objectives; Establishing Performance Standards; Appropriate Practice; Perceived Purpose (Motivation); Evaluation: A Curriculum Rationale (Tyler Model); Defining Content for Objectives; Affective Objectives; Analyzing Learner Outcomes; Knowledge of Results; Teaching Units and Lesson Plans; The Teaching of Reading; Discipline in a Classroom; Modern Measurement Methods; Instructional Supervision; A Criterion-Referenced Strategy; and Experimental Designs for School Research. The 18 programs fall into three general categories--instructional objectives, instructional sequences, and evaluation.

II. Analysis of Classroom Interaction

A. Title: Classroom Behavior Analysis and Treatment

Developer: Robert L. Spaulding

Source: Dr. Robert L. Spaulding
Director, Child Development Institute
San Jose State University
San Jose, California 95114

Cost: $3.50 plus postage

Materials: Book

Description:

This book reports on a comprehensive system for measuring transactional behavior of children and teachers in classroom settings. It can help teachers to apply information about behavior modification techniques by providing them with a means of assessing behavioral styles and behavioral changes. Chapter titles include: The Coping Analysis Schedule for Educational Settings (CASES); The Spaulding Teacher Activity Rating Schedule (STARS); Using CASES and STARS Together to Measure Transaction; Using CASES or STARS in Teacher Training; Using CASES Categories to Measure Change toward Behavioral Objectives; Designing a School Management System to Permit Treatment Schedules by CASES Style; Data Gathering Procedures Using CASES and STARS.

B. Title: Discovering New Dimensions in the Teaching Process

Author or Developer: Greta Morine, Robert Spaulding, and Selma Greenberg

Source: College Division
Intext Educational Publishers
Scranton, Pennsylvania 18515

Cost: $3.95

Materials: Textbook

Description:

This book is designed to help teachers analyze and evaluate their verbal interaction with students in terms of its relationship to their instructional objectives. It focuses on how certain techniques, such as presenting information, asking questions, and reacting to student responses, might vary according to different instructional roles, such as "intellectual authority" or "intellectual guide." Practice exercises are based on transcripts of actual dialogue from both secondary and elementary school classrooms. Immediate feedback (correct answers) follows each practice exercise. Chapter titles include: Categorizing Teaching Activities; Recording Teaching Activities; Analyzing Recorded Behavior; Patterns of Interaction; Evaluating Teaching Behavior; Categorizing Teacher Questions; Teacher Roles and Teaching Behavior.

C. Title: Guided Self-Analysis: Teaching for Inquiry

Author or Developer: Theodore Parsons

Source: Guided Self-Analysis Professional Development Systems

2140 Shattuck
Berkeley, California 94704

Cost: $50.00 for set of six manuals

Materials: Written materials; audio- or videotape lessons and facilities are required but not provided by the program

Description:

The program aims to improve instructional competence by having the teacher learn a series of focused interaction codes, a system for computing profiles for each code, and guidelines for interpreting the profiles. He then codes, computes, and analyzes his own teaching. The program has six sequential coding schedules. Each schedule focuses on a specific type of teaching behavior, e.g., teacher question, teacher responses, type of teacher talk, etc. However, the schedules are interrelated and the interpretation of teaching style and classroom interaction is based upon the relationships among the codes.

The six schedules (focused codes) are: Schedule A, Questioning Strategies, classified according to the type of thinking required of the pupils; Schedule B, Response Patterns, classified according to whether they promote or inhibit further pupil thinking; Schedule C, Teacher Talk Patterns, a breakdown of the time spent in questions and responses, instruction, classroom management, etc.; Schedule D, Teacher-Pupil Talk Patterns; Schedule E, Experience Referents--analyzes extent to which teachers' questions relate to pupil experiences; and Schedule F, Levels of Thinking--analyzes congruence between level of thinking required by teacher questions and level represented by pupil responses.

D. Title: Interaction Analysis

Author or Developer: John H. Hansen and Robert A. Anderson

Source: Northwest Regional Educational Laboratory
Communications and Disseminations
710 S.W. Second Avenue
500 Lindsay Building
Portland, Oregon 97204

Special Ordering Instructions:

Trainer's manual includes copies of student materials and transparencies, which can be duplicated locally. A training tape and training film are obtainable from:

Teacher Inservice Programs and Services
P.O. Box 465
Eugene, Oregon 97401

A filmstrip with audiotape is obtainable from:

University of Minnesota
Audio-Visual Center
Minneapolis, Minnesota

A textbook (Interaction Analysis: Theory, Research, and Application, by Amidon and Hough) is obtainable from:

Addison-Wesley
Reading, Massachusetts

Cost:	Trainer's Manual	$ 4.00	
	Training Tape	12.00	
	Training Film	5.00	rental, or
		225.00	purchase
	Filmstrip with audiotape	20.00	
	Textbook (paperback)	4.50	

Materials: Printed materials include a trainer's manual, a textbook, and handout materials; a training tape, film, and filmstrip with audiotape are necessary supplements to these materials.

Description:

This course in interaction analysis is designed as a 40-hour workshop in use of the Flanders-Amidon matrix. It deals with category definitions, tallying in the matrix, analysis of the matrix, categorical teaching, effect of categories.

E. Title: Minicourse on Interaction Analysis

Author or Developer: Far West Laboratory for Educational Research and Development
1855 Folsom Street
San Francisco, California 94103
(415) 565-3016
Attn: Mrs. R. A. Elder, Ext. 74

Source: Not yet decided; scheduled for January 1973

Cost: Not yet determined, but less than $40.00

Materials: Teacher's manual, answer book; supervisor's manual; six cassette, magnetic voice recordings; orientation package

Objectives:

1. Trainees will learn to encode and decode with Flanders I.A. categories including time-line displays.
2. Trainee will practice analyzing his balance of teacher initiation and response.
3. Trainee will practice producing "because extensions."
4. Trainee will practice accepting and using student ideas and feelings.
5. Trainee will design and conduct a self-inquiry project.

Teachers can work singly or in groups. Two or more persons are recommended.

Availability: The operational field test edition (not final) is available by writing to the Far West Laboratory.

F. Title: Learning Interaction Analysis: A Programmed Approach

Author or Developer: Miles C. Olson

Source: Educational Consulting Association, Inc.
3311 South Broadway, Suite 304
Englewood, Colorado 80110

Cost: $4.95 each (includes book and records)

Materials: A 45-page booklet; two 33-1/3 RPM records

Description:

A compact, entirely self-instructional programmed course designed to introduce students to the concept of interaction analysis and develop their knowledge and coding proficiency using the Flanders Category system. The booklet is divided into four parts: (1) material describing interaction analysis and the Flanders Category system; (2) questions and exercises to learn and practice the category system; (3) material and exercises on completing a matrix; and (4) information on how to interpret the matrix.

The material is carefully sequenced to develop first the student's knowledge of the category system and then his coding skills for both the category and the unit of analysis (time).

G. Title: Systematic and Objective Analysis of Instruction

Author or Developer: James R. Hale and R. Allan Spanjer

Source: Northwest Regional Educational Laboratory
710 S.W. Second Avenue
500 Lindsay Building
Portland, Oregon 97204

Cost: Information not available

Materials: Printed trainer's manual; printed participant materials; audiotape, videotape, or film of classroom teaching episodes to be furnished by user of this program

Description:

This program is designed as a four-week workshop. It has two major phases. The first deals with developing effective interpersonal communication skills. Topics include: group processes, communication skills, constructive use of feelings, interpersonal effect of various responses, processes of interpersonal influence. The second phase deals with developing effective supervision skills. Topics include: a model for systematically improving instruction, systematic analysis of teaching performance, strategy and the conference, problems and issues in improving instruction and supervision.

H. Title: Systematic Observation of Teaching

Author or Developer: Richard L. Ober, Ernest L. Bentley, and Edith Miller

Source: Prentice-Hall, Inc.
Englewood Cliffs, New Jersey

Description:

This book trains teachers to use the Flanders Interaction Analysis system to observe and control their own behavior. It is replete with examples of classroom interaction.

I. Title: The Role of the Teacher in the Classroom

Author or Developer: E. J. Amidon and N. A. Flanders

Source: Paul S. Amidon and Associates
5408 Chicago Avenue South
Minneapolis, Minnesota 55417

Cost: $1.50 (approx.) per copy

Description:

This is a manual written especially for classroom teachers. It introduces the notion of investigating verbal communication as an approach to the improvement of classroom instruction. Methods of interpreting a matrix are explained.

A variety of supportive instructional aids can be obtained from the publisher. These include voice tape recordings with a transcript for practicing coding with the basic 10-category system, videotapes for the same purpose, and programmed materials which introduce increasingly complex teaching skills. A complete inventory of materials can be secured from the publisher. Consultant services can also be arranged in the design and conduct of workshops for the inservice training of teachers or administrators.

III. Attitude Inventories

A. Title: Diagnosing Classroom Learning Environment

Author or Developer: Robert Fox, Margaret B. Luszki, and Richard Schmuck

Source: Science Research Associates, Inc.
259 East Erie Street
Chicago, Illinois 60611

Cost: Contact source

Description:

This pamphlet contains 23 diagnostic tools to assess classroom learning climate, social relations, pupil norms, pupil-teacher interaction, pupil self-concept, and parental influences on school adjustment. It also contains suggestions on organizing the data so that the teacher can focus his efforts to change the classroom learning environment.

B. Title: Educator Image Questionnaires

Author or Developer: Educator Feedback Center

Source: Educator Feedback Center
School of Education
Western Michigan University
Kalamazoo, Michigan

Cost: $10

Description:

This organization will send questionnaires for distribution to teachers or pupils, will collect and analyze the responses, and will prepare a confidential profile for the principal or teacher of perceived strengths and weaknesses, complete with suggestions for strategies to improve leadership or teaching effectiveness. A follow-up questionnaire can be used at a later time and the principal or teacher can receive a new profile showing comparative pretest and posttest responses.

C. Minnesota Student Attitude Inventory (grade 7 and up)
Michigan Pupil Attitude Inventory (grades 3-8)

Author or Developer: Ned A. Flanders
Far West Laboratory for Educational Research and Development
1855 Folsom Street
San Francisco, California 94103
(415) 565-3016

Source: Copies available from author

Cost: No cost, user duplicates own copies

Materials: One test booklet; one answer sheet for each test

Objectives:

1. The class average score on these inventories is associated with classroom interaction variables when used for between-classroom comparisons in a sample of similar classes.
2. Test results can be used to compare the same class over time.

IV. Teaching Strategies - Cognitive Orientation

A. Title: Classroom Questions: What Kinds?

Author or Developer: Norris M. Sanders

Source: Harper and Row
49 East 33rd Street
New York, New York 10016

Cost: Consult source

Description:

This book identifies seven cognitive levels of questions based on Bloom's taxonomy and provides exercises in developing questions at various levels for use in the classroom.

B. Title: Development of Higher Level Thinking Abilities

Author or Developer: John A. McCollum and Rose Marie David

Source: Northwest Regional Educational Laboratory
710 S.W. Second Avenue
500 Lindsay Building
Portland, Oregon 97204

Cost: Information not available

Materials: Printed leader's guide; printed participant materials (can be duplicated locally); 30-minute demonstration film; 5 audiotapes; 1 slide tape

Description:

This is a 40-hour course based on Hilda Taba's work on teaching strategies which increase children's skills in categorizing, generalizing, and applying generalizations. This course can be presented as a workshop, extension course, or methods course. The major topics dealt with are: rationale for change, rationale for curriculum development, question strategies and discussion skills, concept diagnosis, interpretation of data, application of knowledge.

C. Title: Facilitating Inquiry in the Classroom

Author or Developer: Fred E. Newton
Northwest Regional Educational Laboratory

Source: Copy-Print Centers
1208 S.W. Jefferson Street
Portland, Oregon 97201
(Leader's guide and participant materials)

Special Ordering Instructions:

Confrontation tapes (audio) can be obtained from:

Rex Recording Studios
931 S.W. King Street
Portland, Oregon 97205

Demonstration equipment for inquiry exercises (pulse glass and bimetallic strip) can be obtained from:

School Teaching Aids and Supplies
1225 Eighth Street
Berkeley, California 94710

Cost:	Leader's guide	$14.00
	Participants' materials	2.50 each
	Confrontation tapes	12.27 set of 4
	Demonstration equipment:	
	Pulse glass	3.00 each
	Strip	12.00 each

Materials: Printed materials include the leader's guide and participant materials; four audiotapes and several items of demonstration equipment are necessary support materials.

Description:

This course in facilitating inquiry is based upon Richard Suchman's inquiry training technique. It is designed as a 40- to 45-hour workshop. The leader is provided with detailed instructions in the leader's guide, and need not be specially trained for this role. The sessions (or subsets) include such titles as: Experiencing Inquiry as an Inquirer; Identifying Possible Risks and Advantages; Allowing Inquiry to Happen; Inquirer Behavior; Practicing Allowing Inquiry Moves; Evaluating Teamwork Relationships; Developing Problem Focuses; Facilitating Growth Moves; Using Tuning-In Moves; Assessing Practice Tapes.

D. Title: Minicourse 1: Effective Questioning: Elementary Level

Author or Developer: Far West Laboratory for Educational Research and Development

Source: Macmillan Educational Services, Inc.
Front and Brown Streets
Riverside, New Jersey 08075

Cost: $1,435 purchase; $210 six-week rental

Materials: Written materials (teacher and coordinator handbooks); filmed materials

Description:

Teachers who take this course learn to increase the effectiveness of discussion and questioning skills by increasing pupils' readiness to respond, decreasing the amount of teacher talk, probing, prompting, and maintaining a flow of discussion.

E. Title: Minicourse 2: Developing Children's Oral Language

Author or Developer: Far West Laboratory for Educational Research and Development

Source: Macmillan Educational Services
Front and Brown Streets
Riverside, New Jersey 08075

Cost: $1,320 purchase; $195 six-week rental

Materials: Written materials (teacher and coordinator manuals); filmed lessons

Description:

In this course for all who work with K-4 children, adults learn teaching skills that expand the language and thought of the child: expanding a phrase to a sentence, modeling new language patterns, modeling positional words in context with objects, modeling action words, etc.

F. Title: Minicourse 9: Higher Cognitive Questioning

Author or Developer: Far West Laboratory for Educational Research and Development

Source: Macmillan Educational Services, Inc.
Front and Brown Streets
Riverside, New Jersey 08075

Cost: $1,080 purchase; $165 six-week rental

Materials: Written materials (teacher and coordinator manuals); filmed lessons

Description:

Most questions asked by teachers require students simply to regurgitate facts. This minicourse helps teachers in intermediate and junior high grades to develop skills in asking questions that lead students to make inferences and judgments, to solve problems, and to make predictions.

G. Title: 1. The Basic Course: Teaching Is Learning to Listen

2. Activities in Metaphor

Author or Developer: W. J. J. Gordon and T. Poze

Source: Synectics Education Systems
121 Brattle Street
Cambridge, Massachusetts 02138

Costs: $50.00

Materials: Two paperback books

Description:

The two books can be used independently or in conjunction with one another. The major emphasis in both books is the development of the teacher's ability to use metaphoric activity in application with learners. The materials are designed to be self-instructional.

1. "The Basic Course: Teaching Is Learning to Listen" is described as a programmed teacher training course. The objective of the course is the development of teaching skill for the utilization of metaphoric activity for creativity and problem solving. The teacher "learning to listen" refers to the acceptance of learners' feelings and ideas, an attitude of caring about the learner, and the ability to understand the thinking processes and kinds of connections necessary for metaphoric activity. There is a short introduction to synectics theory followed by practice exercises and problem-solving units for teachers and learners.

2. "Activities In Metaphor" is a compilation of activities for students based on metaphoric activity developed in the basic course. These activities are game-like and are developed in step-by-step detail. All of the activities are developed in terms of goals and processes. Examples of activities and their labels are: Answer Sharing, Metaphorically Based Twenty Questions, Matching Pictures, Helen Keller, Charades Based on an Analogy, The Fly's Eye, and The Way It Might Have Happened. Some of the activities lead to others: for example, the charades game may lead to theater exercises, improvisation, and dramatic production. Many of the activities are connected to specific content areas such as art, music, science and mathematics.

H. Title: Three Teaching Strategies for the Social Studies

Author or Developer: Bruce Joyce, Marsha Weil, and Rhoada Wald

Source: Science Research Associates
259 East Erie Street
Chicago, Illinois 60611

Cost: $ 3.60 per manual
$50.00, 3 filmstrips and audiotapes

Description:

The manual explains three teaching strategies: group investigation, concept learning, and role-playing. The filmstrips and audiotapes illustrate these strategies in use in classrooms. The materials train teachers to use these three strategies in their own classrooms.

V. Teaching Strategies - Affective Orientation

A. Title: Human Relations Unit

Author or Developer: Far West Laboratory for Educational Research and Development

Source: Anti-Defamation League
315 Lexington Avenue
New York, New York 10016

Cost: $410 sale; $50 rental

Materials: Films; handbooks

Description:

Four films depict critical encounters, simulated by teachers and students, in an open-ended format that is often based on a question (e.g., "What do you do if a defiant student refuses to remove his sunglasses in class?") After watching each episode, trainees split into small groups to discuss the issues raised, relate them to experiences in their own schools, and determine specific ways to deal with the problems.

The black-and-white 16mm films focus on four related topics: (1) how visitors to a school may be alienated by the attitude and behavior of the school personnel; (2) how a teacher's words and expressions may be offensive to the students; (3) how a teacher can mishandle rules and regulations; and (4) how a teacher's method of dealing with civil rights issues in a classroom discussion can result in confrontation situations.

B. Title: Interpersonal Communications

Author or Developer: John Wallen

Source: Northwest Regional Educational Laboratory
710 S.W. Second Avenue
500 Lindsay Building
Portland, Oregon 97204

Cost: Information not available

Materials: Printed instructor's manual; printed participant materials; 9 films; 1 audiotape

General Description:

This program is designed for use in a workshop setting. Materials are based on the work of the National Training Laboratories, Institute for Applied Behavioral Science. The following topics related to interpersonal communications are dealt with: paraphrasing, behavior description, describing feelings, nonverbal communication and perception check, the concept of feedback, expectations and communication, the interpersonal gap, the effects of feelings, matching behavior with intentions, open communications, communicating about interpersonal relationships, roles and patterns of interpersonal communications, norms and communication, one- and two-way communication, communication patterns in the school building, communicating under pressure, assessment of knowledge, improving skills, developing support for continuous learning.

C. Title: Teaching Achievement Motivation

Author or Developer: Alfred S. Alschuler, Diane Tabor and James McIntyre

Source: Education Ventures, Inc.
209 Court Street
Middletown, Connecticut 06457

Cost:	Text alone	$4.95
	Sampler set (text and pamphlets)	7.50

Materials: Paperback manual and related pamphlets (selected primary source materials available from a general library collection)

Description:

Achievement motivation training is concerned with increasing students' need to achieve their own kind of excellence. "Need to achieve" involves three elements planning to attain excellence, strong feelings about doing well, and action strategies. A repertoire of procedures (techniques) is used to develop need achievement. These include imagination exercises, action exercises such as role plays, games, exercises which emphasize emotional responses, and procedures which emphasize the "here-and-now."

The manual is divided into four parts: backgrou in achievement motivation and psychological education, an achievement motivation workshop for teachers, motivation in classrooms, and achievement motivation training for students. The heart of the course is the teacher-run workshop (Chapter Two). It is designed as a ten-session (approximately 30-hour) course whose goals are to increase knowledge about need achievement and psychological education methods for increasing motives, and to apply psychological education. Participants in the course learn about their own motive patterns and relate their experience to their teaching. Some sessions are optional.

VI. Classroom Organization

A. Title: Team-Teaching Modules

Author or Developer: L. Jean York
Research and Development Center
for Teacher Education
University of Texas
Austin, Texas

Source: The Leslie Press
111 Leslie Street
Dallas, Texas 75207

Cost: $19.60, set of seven volumes
(excludes nonwritten materials)

Materials: Written materials; films; filmstrips; videotape

Special Ordering Instructions:

Ordering instructions and costs for films and other multimedia materials are listed at the beginning of each module.

Description:

This series of modules has as its general objectives: (1) development of an understanding of the philosophy, background, and processes of team teaching --the knowledge of the kinds of personnel and roles composing a teaching team, team-planning, nongraded school organization, and personal characteristics of team members; (2) knowledge of the organizational, diagnostic, and evaluative techniques for grouping and for individualizing instruction; and (3) personal experience in self-paced, individualized instruction and self-evaluation. The series consists of seven modules:

I. The Background, Philosophy, and Purposes of Team Teaching

II. The Roles of the Professional and Paraprofessional in Team Teaching

III. Materials and Resources Needed for Team Teaching and Individualized Instruction

IV. Grouping Children for Instruction in Team Teaching

V Team Teaching as a Facilitator of the Nongraded School

VI. Evaluation of Team Teaching and Children's Continuous Progress

VII. Prerequisites for Good Planning Session in Team Teaching

B. Title: Minicourse 5: Individualizing Instruction in Mathematics

Author or Developer: Far West Laboratory for Educational Research and Development

Source: Macmillan Educational Services, Inc.
Front and Brown Streets
Riverside, New Jersey 08075

Cost: $1,395 purchase; $195 six-week rental

Materials: Written materials (teacher and coordinator manuals); filmed instructions; models

Description:

This course helps the elementary teacher handle individual instruction. It provides tutoring techniques to improve mathematics skills through diagnosis, demonstration, evaluation, and use of assigned practice examples for estimating, number of operations, verbal problems, etc.

C. Title: Minicourse 8: Organizing Independent Learning: Primary Level

Author or Developer: Far West Laboratory for Educational Research and Development

Source: Macmillan Educational Services, Inc.
Front and Brown Streets
Riverside, New Jersey 08075

Cost: $1,080 purchase; $165 six-week rental

Materials: Written materials (teacher and coordinator manuals); filmed lessons

Description:

During this minicourse, K-3 teachers learn a set of organizational procedures that make it possible for them to work with a small group of children for 15 to 30 minutes while the remaining students carry on independent activities. The children learn how to anticipate and deal with problems, to set their own goals, and to evaluate their progress.

D. Title: Teaching in IPI (Individually Prescribed Instruction) Mathematics

Author or Developer: Research for Better Schools

Source: Research for Better Schools
1700 Market Street
Philadelphia, Pennsylvania

Cost: Six-volume set (includes records), $10.00

Materials: Six workbooks; record set of three 33-1/3 RPM records

Description:

IPI Mathematics is a system of individualized instruction for children which coordinates behavioral objectives, diagnostic instruments, curriculum material, teaching methods, and instructional time (pacing). This self-instructional program, Teaching in IPI Mathematics, is designed to train teachers in the general diagnosing, planning, and guiding skills involved in individualizing instruction, and specifically to implement the IPI Mathematics Curriculum.

Six volumes deal with the following areas: (1) an overview of individualized instruction and IPI (Volume I); (2) behavioral objectives and the specific objectives in IPI Mathematics (Volume II); (3) diagnosing student achievement (administering and prescribing on the basis of student achievement tests - Volume III); (4) developing prescriptions for three case studies (Volumes IV, V, and VI).

VII. Administrative Skills

A. Title: Determining Instructional Purposes

Author or Developer: Far West Laboratory for Educational Research and Development

Source: Far West Laboratory for Educational
Research and Development
Educational Management Program
Hotel Claremont
Berkeley, California

Cost: $6-$17 for each training unit

Description:

These are three self-contained training units designed for use by a group of administrators. The topics covered are identification and analysis of existing and anticipated instructional programs and needs, specification and refinement of desired educational outcomes, and specification of performance standards against which actual outcomes can be measured.

B. Title: (ICE) Individualized Continuing Education for School Administrators

Author or Developer: Charles F. Kettering, Limited
Englewood, Colorado

Description:

In this system, the school district supports the principal in an individually designed program of self-improvement that will contribute to desired improvements in the school and district. Several administrators within a district work together as a study group, as well as participating in individualized learning activities. The superintendent and outside consultants provide program leadership. Districts that have participated in this ICE program through the auspices of the Kettering Foundation have demonstrated that a significant program can be operated for approximately $1,000 per year per participant. Thus, in a district with 24 principals, each principal could participate in a self-renewal program every fourth year, with an annual cost of $6,000 to the district for the overall program.

C. Title: Research Utilizing Problem Solving (RUPS)

Author or Developer: Charles Jung, Rene Pine, and Ruth Emory

Source: Copy-Print Centers
1206 S.W. Jefferson Street
Portland, Oregon 97201

Special Ordering Instructions:

Tape recording obtainable from:

Rex Recording Studios
931 S.W. King Street
Portland, Oregon 97205

Cost:	Leader's Manual	$6.00
	Participant Materials	2.25
	Tape Recording	3.25

Materials: Printed manuals for use by leader and students; tape recording; printed charts to be prepared by leader

Description:

A carefully designed and tested course requiring 35 hours of instruction, designed to develop skills in problem solving. These skills include: identifying the problem: analysis of the problem using the force field technique (Lewin); selecting tools for data collection; deriving implications and action alternatives from research findings; planning for action.

D. Title: Self Performance Achievement Record (SPAR)

Author or Developer: James L. Olivero et al.
Program Associates
Charles F. Kettering, Limited
Englewood, Colorado

Description:

In this program, the principal's self-development plans are laid out in a contract, with specification of behavioral objectives and acceptable evidence that performance criteria have been met. Forms and instructions to help the principal plan his own inservice education have been developed by CFK Ltd. Associates.

REMARKS BY GRETA MORINE

I want to make a few comments about the relationships I saw between the papers that were presented earlier and the paper that Dr. Flanders and I wrote. Dr. Glass recommended at the end of his paper an observational/judgmental system of evaluation, and Henry Kaiser in his discussion of that paper pointed out the expense of using this kind of method. The audience raised questions about methods of training personnel to carry out an observational evaluation system and about the possibility of utilizing methods of self-evaluation. What we have tried to do is to recommend three possible systems of assessment--the use of inquiry projects, the use of competency contracts, and the use of directed skill training. We might borrow Dr. Eisner's phrase and call this "personalized assessment," since the suggestion is that a teacher might select the one system that is most suitable for him. None of these three systems, we think, would be prohibitively expensive. The appendix of our paper gives the cost of the materials listed there. They were selected in part because they are self-instructional and in part because they are available now at a reasonable cost. Two of the three systems we set forth involve a form of self-evaluation.

Dr. Popham in his paper differentiated between process and product. He indicated he was most concerned with the assessment of product in dealing with the section of the Act that related to pupil achievement. Our topic, it seems to me, is more closely related to process, but I think it would be fair to say that we have considered both process and product in the paper. In identifying teaching behaviors to be assessed, we have selected those that have been shown to be related to pupil outcomes, both achievement and attitudinal. Further, we have selected assessment processes to recommend on the basis of the assessment product we consider to be most important--the improvement of instruction.

Dr. Eisner, in his reaction to Dr. Popham's paper, made the suggestion rather strongly that accountability, evaluation, and measurement could not rest on prediction. It seems to me that the inquiry projects we have suggested are one solution to that problem--in other words, that we can base assessment on things other than prediction.

Dr. Abbott in his paper defined accountability as holding someone responsible for accomplishing something that he has agreed to accomplish. Our recommendations for the use of inquiry projects and competency contracts are ways of implementing that definition of accountability. He suggested that one way to assess administrative competency was to measure employee morale. We've tried to identify additional criteria that might be used; for example, the number of successful inquiry projects teachers are able to carry out, the number of competency contracts teachers have completed, and the number of teachers who engage in directed skill training and thus improve their instructional competency.

To summarize briefly, there are relationships between the problems raised in earlier papers and discussions and proposals made in our paper. We do not pretend to have a solution for each of these problems. But we have addressed ourselves to them.

DISCUSSION OF PAPER BY FLANDERS AND MORINE

JOHN YOUNG
Mount Diablo (California) Unified Education Association

I was impressed, in reading through the paper by Dr. Flanders and Dr. Morine, with the feeling that its general direction seems to be toward the improvement of instruction. I am very pleased to note that, at least in that paper and hopefully in all of the discussion that has taken place thus far in this conference, the emphasis has been on how we can improve instructional programs. I hope we can use the Stull Act for that purpose rather than--what has come to be a matter of great fear to many teachers in California--as a way to get rid of teachers. I think this would be an unfortunate consequence of what is potentially an opportunity to do some real good. In my opinion no one really knows yet the consequences of the Stull Act, but if we use it properly I think it can do a great deal of good.

I have made several notes concerning the paper by Drs. Flanders and Morine. On balance, I think, the paper can be looked upon as a very positive contribution. I do feel, however, that far too little attention has been given in the paper to the evaluation of nonteaching certificated personnel. It is true that attention has been paid to the fact that the Stull Act requires the evaluation of each certificated person, which includes such nonteaching people as administrative and supervisory personnel. But there are many references throughout the paper to the teacher and his supervisor in the evaluation process. I would have felt more comfortable, personally, had the paper talked about the evaluator and the evaluatee, thus letting it be recognized that there are indeed people with varying functions within the school system.

I find some difficulty in accepting the authors' definition of adjunctive duties, when they say that "adjunctive duties are presumably to be distinguished from active teaching, administrating, or supervising." I have some belief that it would be better to say that adjunctive duties are those which are not directly related to the primary, or basic, duties. A very good point made in the paper is that adjunctive duties will include such matters as the need for the evaluatee (my language) or the teacher or administrator to evaluate his own performance. I certainly agree this is an important responsibility, but I don't think this is an adjunctive duty--I think it is a basic requirement, part and parcel of the evaluatee's primary responsibility.

A large part of the paper was devoted to the evaluation process as it applies to the teaching act and the various aspects of teaching, and the fact that administrators and supervisory personnel are also to be evaluated is rather incidental. That is regrettable, I think, but on the other hand there is another way to look at this, and that is the fact that we spend more time talking about evaluating teaching because teaching is the most important job to be done. Therefore, of course, we should spend more time talking about the evaluation of teaching.

Also, there appears to me to be a mistake in the interpretation of what the Stull Act indeed does require in Article 5.5, dealing with evaluation. I quote from the paper. "Competence and standards are probably used in the language of the Act in the simple sense of performing one's job in a satisfactory manner." Now, I really don't think that is right. The Act clearly talks about standards of expected pupil progress. Without arguing whether or not that's a good thing, it seems to me that at least the language of the law talks about standards of expected pupil progress and then by implication, I suppose, one could argue that we must consider the performance of the teacher or the administrator in terms of the kind of performance they secure from students. It is precisely at this point, I think, that so many teachers are very concerned about the possible consequences of this Act.

There is a growing reluctance throughout California to rely upon any kind of standardized tests. Indeed, we were very successful this year in getting the state legislature to make it very clear in a piece of legislation that standardized tests are not going to be the vehicle by which we determine whether the standards of expected student progress are realized. This was a unanimous position of the legislature, but there was one important individual we failed to convince in the State, and that was the Governor, who vetoed the measure.

I am pleased that Dr. Flanders and Dr. Morine have flexible goals which are set by the evaluatee. One of the large goals referred to here is the improvement of attitudes of students. I believe that both teachers and administrators have a great opportunity to do something about setting the climate for good instruction and that it is important to talk about the desired end products rather than just about a process, with one of the end products being the improvement of student attitudes. This is a very positive thing; I am very pleased to have it stated and also to have it indicated that there are measurement devices available to help achieve this goal.

A problem recognized by many of us who have spent many years in the classroom is that although academic studies are very valuable, they have a great potential for improvement. It is too easy to disregard such studies because so often there is not sufficient time to deal with things which we might regard as somewhat esoteric and unattainable. I think that's unfortunate; somehow we must find a way to convince our employers, our boards of education, to endow the educational program with enough resources so that we can bring about the improvement we all want. This will take some thought and planning, of course. It will require some hard work based on measurement results and on careful evaluation of what is going on. To implement the good suggestions put forth in this paper and at this conference will take a job of selling at the district level--persuading the people who are the decision makers to provide the time and resources needed by teachers and administrators to do the job.

I would like to remark on one exceedingly good comment, relative to the required amount of classroom observation, in the paper we have just heard. I think Dr. Flanders indicated that at least six observations of a teacher in the classroom would be necessary in order to have any kind of an adequate evaluation of classroom performance. Let us say we are considering an elementary school with one principal and 30 teachers, with six one-hour visits to each classroom. Then let us consider the out-of-class time the evaluator must spend on evaluating what he has seen and in holding conferences with the evaluatees. The administrator/evaluator will thus spend four or five hours a week on evaluation alone, each week of the year. Maybe that is a reasonable amount of time, but it may necessitate some reordering of priorities. I suggest to the administrative people here that they may need to start deciding what they are doing right now that they can off-load, or stop doing, because there isn't going to be time to do everything.

Finally, throughout this paper and anything else in the literature I have read, there is not much reference to the use of teachers in the evaluation of administrative and supervisory personnel. Now either a teacher is a professional or he is not. It is one thing to give lip service to this concept; it is another thing to make it true. What is a colleague? I see a reference in this paper to the teacher and his colleagues, which raises the question in my mind, do we look upon the teacher and the administrator as colleagues? I am troubled, frankly, by the definition frequently used of a "peer." Is a teacher a peer of an administrator or isn't he? I claim that he is. On the other hand, if the size of their paychecks is the standard by which we make the distinction, then perhaps they are not peers. But I would like to leave you with the thought that any evaluation program in any district which does not provide the opportunity for teachers to be plugged into the evaluation process in both ways, not only evaluating other teachers but also evaluating the person who is supposed to support them in the teaching act, is a very inadequate kind of an evaluation policy.

KEITH WAITE
Tulare (California) Union High School District

When our staff first took a look this spring at the mandates of the Stull Act, we were appalled, perhaps chagrined, but mostly overwhelmed. It reminded me of a story told me by a lady who some 60 years ago taught school in a very tough portion of Brooklyn. She had taught a few years, but it was her first year in this school, and she got the hardest classroom to deal with. The method she used to try to survive was to pick the toughest kid of all, who was also a leader, and to ask him if he would help her maintain a proper climate of control and a suitable learning environment, I suppose, and then to assist him so that he didn't do anything that would land them both in jail. This worked out very well and both the teacher and the class had a good year. The young man she had picked to help her also had a beautiful year, of course. At the end of the school year he came to her to thank her for what had happened. He said: "Mrs. Michelle, when I grow up

I am going to make lots of money and I am going to build a brand-new school for you--without a single kid in it!" His concept of teacher heaven is something like our concept of what we need if we are really going to implement the Stull Act. But unfortunately we don't have this utopia, so we have to proceed under other circumstances.

Prior to this conference, frankly, I had not dwelt much upon the aspects of the Stull Act which are the topics of discussion in the paper we have just heard--adjunctive duties, proper control, the suitable learning environment. I was impressed and awed by the paper. It is a good deal more scholarly than the kind of thing I grapple with day by day as I go about my job as a district superintendent. I am impressed by the self-evaluation aspects of the paper, although I have some of the same criticisms mentioned by the preceding speaker. I am not going into these, however, because my real dissatisfactions are not objections to the paper.

I think the meat of the Stull Act is in Parts A and B. I believe that what the nonteaching people in the state--school board members, parents, and legislators--support in this Act is the idea that some time they expect us to tell them what we're trying to teach their kids and whether or not we have accomplished it--and if not, why not. They also expect us in some fair and impartial way to relate our success and our failure in accomplishing this to the competency and the efforts of the staff in charge of the school.

I believe the portions of the Stull Act we are discussing at this session are really concessions by the Legislature to the old order--to the educators. It's like a girl writing a "Dear John" letter to her fiance overseas to explain that she is going to marry the boy who stayed home, who adds at the end of the letter, "P.S. I love you." Under these sections of the Act we are still talking about control, about adjunctive duties, and about a suitable learning environment. We are talking about processes--and that is what we are accustomed to evaluating. We are used to going into a classroom, watching a teacher teach, coming out, and saying, "The teacher did this, the room looked like this," and so forth. But what the Stull Act is asking us to do is to look at pupil progress, and I think this is where we should be putting our attention.

Now, it is perhaps helpful and good to have the section on adjunctive duties, proper control, and suitable learning environment in the Act. We can use these requirements while we are trying to implement the other portions of the Act, which is going to take a long time. There is a danger, however, that this could be a retreat to keep us from looking at the other portions of the Act, which are very complex and perplexing, but occasionally a little bit exciting.

The Act opens up some possibilities that we may be able to look at education as something which is very meaningful. If we really can identify how children learn and if we can substantially agree on what they should learn, then the control and the adjunctive duties and the learning environment problems become part of the process. If children are really learning in someone's classroom, who is going to care about their control as long as they don't tear up the furniture or disturb someone next door? If we really can determine that a teacher, a supervisor, or a superintendent is producing, all these other factors become a part of why it happens.

Please understand, I don't think the paper we have just heard is about unimportant things. I think it is about very important things. It is just that I don't think it is about the central subject of the Stull Act, which is student progress, and that is what we should be concentrating on. If the principals in our district came to me, for example, and said, "We have just read an exciting paper by Flanders and Morine, and in order to assess whether our people have proper control and are maintaining suitable learning environments, we would like to make six visits to each classroom and do the things they suggest doing," I really wouldn't be very happy. I would say to them, "I don't care whether or not you make a single visit to a classroom this year. But I want you to spend whatever time you and the teachers and everyone else can find in developing some description of what we want students to learn and how we are going to assess whether they have learned it. That is what we should spend our time on this year!"

On the other hand, let's suppose they said: "We've read this elaborate paper by these talented people, and some of us --supervisors, principals, and a group of teachers--would like to get involved in interaction analysis and some other in-service learning activities, because we feel that if we can improve in these areas it will improve student progress." Then I would be ecstatic. I would, that is, if it is possible for a tired old superintendent like me to become ecstatic any more. And I would certainly be pleased that they had chosen Drs. Flanders and Morine to take their ideas from.

THE USE OF EVALUATION TO IMPROVE PERFORMANCE

MARVIN C. ALKIN
EVA L. BAKER
University of California, Los Angeles

The public has demanded that schools "be accountable." These demands generally reflect a concern for "proper" utilization of resources, but what is called for is far more than mere financial accounting to ensure that funds are not illegally spent or embezzled. Schools are being asked to demonstrate that the outcomes they are producing are worth the dollar investment provided by communities and that schools have mechanisms for instructional improvement.

The Stull Act is part of this general move toward accountability. It is an attempt by the California Legislature to meet the public demands for accountability by transforming talk into action. Through passage of the Stull Act, the legislature has focused its concern with accountability on evaluation of teachers and other certificated personnel. The Act stipulates that each school district must establish, with the advice of the local teachers' organization, an objective system of evaluation for annual appraisal of probationary teachers and biennial appraisal of all other certificated personnel. The Act further requires that any such system must include established standards and techniques for the assessment of student progress in each area of study, assessment of teacher competence as it relates to established standards, assessment of duties performed by certificated personnel as adjuncts to their regular assignments, and procedures for ascertaining that teachers are maintaining proper control and a suitable learning environment in their classrooms. The results of the evaluation of a certificated employee must be transmitted to the employee in writing and discussed in a meeting between evaluator and evaluatee. Should an employee's professional competency be judged below standard, the school board is then required to assist the individual in upgrading his or her performance.

The Stull Act has been viewed by some as a punitive measure, designed to harass teachers and other school personnel. It is often seen as a threat. However, while the immediate purpose of the Stull Act is to hold teachers responsible for their performance and the performance of their students, its ultimate purpose is to increase the effectiveness of school systems by improving the professional skills of certificated personnel. That is to say, it is hoped that the Stull Act and the current emphasis on accountability in education will result not just in the policing of teachers' activities and performance, but in better teaching, more effective school systems and, ultimately, a better education for each student.

Accountability--What Is It and What Does It Mean?

The foregoing description of the Stull Act is embedded in the context of the general demand for "accountability." But what is meant by this term? Who is accountable to whom? What are they accountable for? Is teacher evaluation synonymous with accountability? Lack of an adequate description is a major problem in dealing with accountability. A person investigating the area for the first time is confronted with a plethora of views, schemes, mechanisms, and definitions that pass under the guise of accountability.

In order to provide an adequate framework within which to discuss the Stull Act, a working definition of accountability must first be established. Accountability is a relationship in which the participants agree in advance to accept specified rewards and costs on the basis of an evaluation of specified ends. The essence of this definition is that a relationship exists in which each of the participants agrees in advance as to the evaluation procedure that will be used to determine acceptable performance. The evaluation procedure includes the measures to be used for demonstrating the attainment of the specified criteria and also the standard or level of attainment required on these measures in order to achieve acceptability. Finally, the negotiants stipulate a set of rewards/penalties that will be attached to compliance/noncompliance.

From this definition four major elements in the accountability relationship can be delineated. These four basic elements may be summarized as follows:

1. Agreement
2. Evaluation
3. Rewards and penalties
4. Public disclosure

At the foundation of any accountability relationship is the concept of agreement. Agreement generally implies the mutual acceptance of a position or a set of conditions. As an element of accountability, agreement refers to the mutual acceptance of the evaluation procedures to be employed, of the tolerable constraints, and of the rewards/penalties that will ensue from compliance/noncompliance. It is rather clear that agreement must be reached with respect to the evaluation procedures and the rewards. Although not so obvious, school participants must also understand the necessity for reaching agreement on allowable constraints. That is, it is important in accountability to specify the nature of the students to be worked with as well as the instructional materials and situations that will be provided. It is impossible to discuss anticipated outcomes without using school and instructional constraints as a frame of reference.

The second element, "evaluation," is a topic of consideration later in this paper and hence will not be discussed at this time. Individuals are not always self-motivating simply on the basis of the stimulus of personal improvement. A system of rewards for attaining stipulated goals and penalties for lack of attainment appears to be a necessary third element of an accountability system. In today's society rewards are likely to take the form of job security and money.

The final element of the accountability relationship is a spirit of disclosure or a feeling that the public has a right to know what is happening. This public access to evaluation data can only be meaningful when grounded in a system in which the components are not only agreed upon but are also explicit and relevant. This last caveat may be clearly understood when one examines the history of public release of standardized test scores on a school-by-school basis without an accompanying explanatory discussion or indication of unique constraints present at each school. Accountability systems must be established in such a way that evidence is presented in a manner particularly relevant to public concerns about the entreprise.

It should be clear that teacher accountability, and in particular the Stull Act in California, falls under the broad definition of accountability that has already been established. The worth of an educational enterprise--teaching--is being demonstrated by assessing student achievement and teacher behavior indices. Yet there is much more to accountability than teacher evaluation. In fact, any attempt to consider only teacher accountability is wholly inadequate because of the nature of the interrelationships between various accountability levels.

Accountability occurs within the multidimensional context of educational institutions. It is simplistic to say that teachers are accountable or they are not without considering at the same time the other interrelated accountability levels.

Conceptually, the consideration of three levels of accountability will be proposed in this paper: goal accountability, program accountability, and outcome accountability. These three accountability levels derive from an attempt at answering the question, "Who is accountable to whom and for what?"

The first area to be considered is goal accountability. School boards are accountable (or should be) to the public for everything that they do. The foundation of this accountability relationship is in educational goals. School boards are accountable to the public for the proper selection of goals. After all, school boards are legally supposed to function as the lay group expressing the desires and wishes of a broader constituency regarding the goals and objectives to be achieved by the educational program. This determination is clearly within the domain of the public's review responsibility. In goal accountability, school boards are accountable to the public for ensuring that the proper goals and objectives are being pursued in the school program.

After goals and objectives are selected, responsibility must be assigned for the selection of instructional strategies deemed most appropriate and effective for achieving the stipulated goals and objectives. This responsibility for program accountability rests generally with school administrators and other members of the professional staff engaged in the program selection-modification-adoption sequence. If teachers are conceived of as program operators (as in the Stull Act), and the intention is to hold them accountable for the outcomes of their activities, then they may only be held accountable within the constraints of the programs with which they have been provided.

Just as a machine operator cannot be held responsible for his products until it has been demonstrated that the machine he has been provided with is capable of producing that outcome, so a teacher cannot be held responsible for achieving learning outcomes unless it has been shown that the program that has been provided to the teacher is capable of achieving those outcomes under proper management. Consequently, those to be answerable for program accountability must not only be responsible for the choice of the program but must also be responsible for validating the program's capability for achieving the desired outcomes. Administrators and other district personnel, though ultimately responsible to the public, are specifically accountable to the school board for maintaining a program appropriate for meeting the stipulated objectives.

There is one further dimension of program accountability to be considered, and that is the stipulation of program capabilities under varying situational constraints. For example, suppose a racetrack driver is provided with an automobile that has a potential speed of 150 miles per hour, validated and verified under optimum driving conditions. Even if he is the top driver available, he probably would be unable to maintain such high speeds under conditions less than optimum, say on rain-slicked roads. Similarly, a teacher, implementing a program under situational constraints different from those under which it was validated, may be unable to achieve the specified outcome standards regardless of the quality of program management. This implies, then, that those to be held responsible for program accountability must stipulate the producible program outcomes in terms of various sets of constraints and the varying inputs that might be encountered. It is necessary to indicate the performance standards for each program in terms of different student inputs and different situational constraints. The extent to which standards, given alternative constraints, can be differentially stipulated has not been clearly demonstrated. Further, the potential political consequence of publicly assenting to different standards for different student groups could be devastating, even when variations in antecedent experiences are clearly documented.

The last type of accountability, outcome accountability, is what is meant, or should be meant, by "teacher accountability" as mandated in the Stull Act. In outcome accountability an instructional leader (usually a teacher) is accountable to the administration for specified pupil outcomes. These outcomes are thought to be a function of his or her management of an instructional program that has certain product capabilities; the task is to determine whether the teacher has managed the program in such a way as to achieve standards or criteria that might be expected from it. In this level of accountability the teacher is not questioned on process characteristics such as a score on a teacher performance test, or the amount of time spent in the classroom, or the processes used. Instead, the teacher is told, "Here is a program whose capabilities have been demonstrated. Show that you are able to produce student outcomes of the desired type and standards using this program."

It has previously been said that teachers may only be held accountable within the constraints of the program with which they have been provided. There are those who would maintain, however, that the accountability concern should not focus upon these constraints since the teacher, to a great extent, is the program. But while one cannot deny that the teacher represents the greatest single factor in program operation, teachers do have to work within constraints, given such input variables as type of students, kind of texts, number of students, etc. Though the effects of some of these constraints may be small, they do affect the management of the program to varying degrees, and to that extent must be considered in outcome accountability.

From the preceding discussion the interdependence of accountability levels becomes evident. Each accountability level is embedded within the previous one and is dependent upon the prior achievement of some form of accountability.

For instance, it does little good to talk about validated programs for unimportant or irrelevant goals. If goal accountability has not be attained and the important goals have not been specified in some systematic way, then how is it even possible to talk about validated programs? Moreover, how can one demand the attainment of outcomes when programs that would make possible such attainment are not available and the constraints of a particular teaching situation make achievement impossible. Thus, as can be noted in Figure 1, outcome accountability is dependent upon the prior accomplishment of program accountability and program accountability is meaningless unless goals have been properly specified.

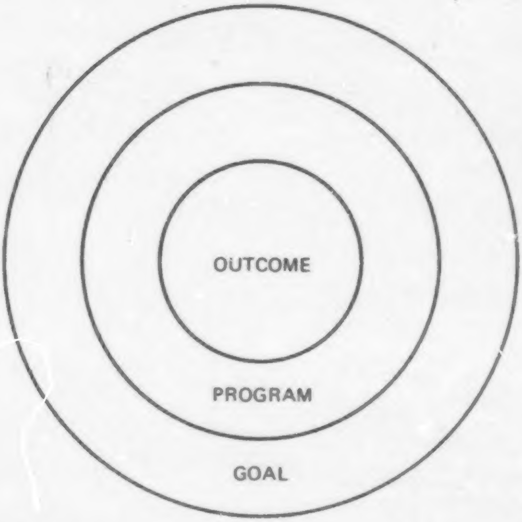

Figure 1.

Theory and Reality

The foregoing discussion on goal, program, and outcome accountability is relatively uncomplicated. This theoretical model reflects an ideal in which various functions are separately compartmentalized. Such compartmentalization helps in understanding more clearly the nature of the various accountability functions and their ramifications. However, linear sequences rarely hold up when set in life conditions, and the neatness of the goal → program → outcome accountability analysis quickly dissipates under real-world pressures. The Stull Act, for instance, has provisions pertinent to both program and outcome accountability simultaneously. The question of goal accountability is largely ignored, under the assumption that goals are someone else's problem (in California, the problem of the Goals Committee).

Successful accountability systems consist of two major components: first, procedures for evaluation, where measurements are made and deficiencies are identified; second, instructional improvement, where remedies for identified deficiencies are hypothesized and implemented. Since functioning accountability systems are regenerative, further measurements and judgments are made following the implementation of an improvement strategy.

The rest of the paper will be directed to practical decisions in the selection of measures and the choice of improvement strategies.

Evaluation and the Stull Act

Evaluation is the beginning and the end of accountability. At both points, bracketing any improvement strategy, a minimum set of deliberate evaluation activities is required. While not unalterably sequential, the following events are suggested as comprising a minimum basis for accountability:

1. Deciding on goals
2. Selecting measures to serve as indicators
3. Setting standards
4. Collecting and analyzing data
5. Identifying deficiencies in performance
6. Communicating deficiencies to relevant decision audiences

To some extent, systematized routines exist that can facilitate the accomplishment of the activities outlined above. In most cases such techniques are in the developmental stage, but in an area such as data analysis the alternatives are familiar and likely to be automatically employed. If we consider each evaluation activity in turn, its relevance for accountability may become more clear.

Deciding on Goals

In any accountability system capable of replication, a statement of goals must be prepared. In one sense, the California Legislature has preempted the goals-setting arena by declaring that improvement of teaching is to be effected; that is the goal. However, operational proximates of that end must be formulated and approved before prerequisite goals are delineated. (In another sense the state mandate that standardized tests must be given in school districts constitutes another preemption in the goal-setting area.)

Needs assessment procedures may be employed to aid in the goal selection and delineation stages. In needs assessment, significant constituencies are queried regarding preferences for alternative educational outcomes. Such preference data are supplemented with reports of status testing undertaken to determine pupil performance ranges in areas of high concern. However, the accumulation of such data does not in itself make the selection decision. The final set of goals will doubtless not only reflect preferences of parents and others about desirable consequences of instruction, but are likely to be constrained by judgments about feasibility and political face-saving as well.

Selection of Measures to Serve as Indicators of Goals

After goals are articulated, measures of goal attainment must be described and then must be selected, adapted, or developed. Often no available measures sufficiently congruent with desired goals will be available. Approximations may be employed, although it is critical that any adaptations in the measures be instituted a priori, so that the specter of data manipulation can be clearly avoided. The local development of measures will rarely be adequate, particularly in light of the time and resource constraints against which California school districts will surely be working. More attention will be devoted to the consideration of measures of teacher evaluation in a subsequent section of the paper.

Setting Standards

In order to discriminate deficient from proficient service, standards of achievement must be set. Setting standards is conceptually an easy task to understand. The question being asked is, "How will we know whether the performance is to be considered satisfactory?" That is, what is the definition of satisfactory performance? Obviously if all students answered every question on every test correctly, there would not be too much question about the standard having been met or exceeded. But surely there must be some lower level that can be agreed upon as being a reasonable standard.

Standards are of several types. On the one hand, internal or "absolute" standards might be established; on the other hand, the standards established might be external or "comparative." In the first case, what is essentially being said is, "These are our goals; this is what we want our students to know, and we will consider that we have been successful if (a stipulated) percentage of our students can attain that goal." For example, a swimming instructor might establish a standard that 90% of his students should be able to swim 100 yards using any stroke or combination of strokes; here he is not concerned with how his students compare with those of other instructors. The standard does not take into consideration the relative status of his pupils vis-a-vis other instructor's pupils; it is absolute, internal, and unrationalized.

Alternately, standards can be directed externally and be comparative in nature. This is expressed in the age-old question that is frequently asked by school board members, parents, and newspapers everywhere: "How are our kids doing compared with kids in other districts?" Or, "How are Mrs. Jones's students doing?" Of course the response to such a question is yet another question, namely, "With whom do you want them compared?" Should comparisons be made with other districts in the state? Or with other districts in the region? Or with other teachers in the district? No matter what standard is set there is the risk of making judgments based on faulty comparison groups. Determining the appropriate group on which standards can be based is a seemingly impossible task. One possibility is to compare achievement with that attained in similar situations. This technique involves predicting an appropriate standard for each teacher based upon the unique inputs of his or her situation. This procedure has been discussed elsewhere (Klein and Alkin, 1972).

Because the technology of criterion setting is weak, the uncertainty high, and the likelihood of error rampant, alternative sets of criterion ranges should be described in initial implementation of accountability models. Lower tolerances of such criteria could reflect the level of present practice and would be nominally in compliance with legislation. Functional criterion setting of this sort should reduce the cataclysms resulting from serious miscalculations, where a spate of corresponding improvement programs would be implied by large numbers of obtained "deficient" areas. While using existing practice as a baseline criterion is palatable when performance levels are seen or felt to be satisfactory, as in certain suburban settings, it is possible that differential criteria will be necessary in schools that are far from "normal."

Collecting and Analyzing Data

Once standards have been set, information that will facilitate comparison of achieved and desired levels of performance is needed. For instance, it must be decided whether all students or only a representative sample should be tested. Although this question is best considered within the context of each evaluation and its purposes, techniques of matrix sampling are available and of great value in many instances (Husek and Sirotnik, 1968). Thus, if the concern is to make evaluative judgments about a teacher based on the performance of his or her students on a large number of criteria, it may be totally inappropriate to give each child the same set of test items since information on how each individual child did on every goal is not needed. Procedures for sampling tests and children would provide accurate data descriptive of teacher performance at a fraction of the cost of total testing.

When the information on student performance has been gathered, the test scores can then be analyzed. Data analysis procedures in accountability concerns follow in the familiar path of most empirical work. In addition, procedures for summarizing data, other than the usual descriptive statistics, may be suitable. For instance, frequency distributions displaying the number or percentage of students who meet given proficiency levels may be more meaningfully interpreted at the classroom or school level than means, standard deviations, or regression residuals.

Identifying Deficiencies in Performance

The task of identifying deficiencies seems almost preordained by the prior activities of selecting measures, establishing standards, and analyzing data. If the establishment of standards is considered in broad terms to include not only the determination of the desired outcomes but also the tolerable ranges of deviation from the desired attainments, the question of what constitutes a deficiency is prejudged.

A more difficult issue, however, is the nature of the process deficiencies. That is, while the diagnosis of outcome deficiencies is reasonably well determined by the previous steps, the determination of the appropriate remediation is not. Unfortunately the research basis for determining appropriate remediation is currently lacking.

Another problem related to the identification of deficiencies pertains to the relative weight of different areas in instances of use of multiple criteria. Do instances of superlative achievement outweigh or compensate for deficiencies in other areas? If so, how many units of mathematics "overachievement" will compensate for how many units of reading "underachievement?" Or how many units of outcome attainment beyond that which was anticipated will compensate for a number of units of lack of "maintenance of adequate control?"

Simple decision models could be devised to cover such circumstances, but it is doubtful whether sufficient thought has been given to enable assigning weights to multiple criteria in judging total deficiency. Nor is it entirely clear whether the legislation of the Stull Act would allow for the construction of such decision models. The legislature seems to have framed the Stull Act in "clear" terms of "success" or "failure," unfortunately ignoring the likely event of differential attainment of multiple criteria.

Communicating Deficiencies to Relevant Decision Audiences

The influence of evaluation results on instructional improvement may in part be related to the audience to whom the results are reported. The importance of framing the decision context as a part of the evaluation activity has been discussed elsewhere (Alkin, 1973). That is, the nature of the data to be selected, collected, analyzed, and reported is dependent upon the characteristics and position of the particular decision maker and the decision purposes for which the evaluation is intended. Thus, the key questions are what audiences are to receive the evaluation data, and what is the potential decision or set of decisions that may occur as a consequence of those evaluation data? This issue really is at the crux of the accountability question in the Stull Act. When the concern is with accountability and the evaluation of teachers, the key question is, who will receive the data and what actions are likely to occur?

The law clearly stipulates that one recipient of this information is to be the teacher himself or herself. Indeed, the format for this data presentation is rather clearly prescribed:

> 1. The evaluatee is provided in writing with a copy of the formal evaluation report not later than 60 days before the end of the school year during which the evaluation takes place.
>
> 2. A follow-up meeting of the evaluator and evaluatee is held to discuss the evaluation report. . .
>
> 3. The evaluatee is provided with procedures to

follow in order to initiate a written response or
reaction to the evaluation.

Clearly, the teacher is not the decision maker in his own evaluation. In the accountability scheme described in this paper an agreement is made between the teacher and supervisor on the outcome levels that will be expected, on the procedure for evaluation, and on the rewards (or punishments) that will be invoked on the basis of the evaluation findings. The teacher will not generally make the decision that he is to be fired or, in potential future modifications of the system, that he is to receive a merit increase in pay. The purpose of reporting evaluation findings to teachers is to provide them with internal feedback in order to facilitate subsequent instructional improvement. This improvement function is different from the teacher evaluation function.

In the teacher evaluation activities of the Stull Act, the question must inevitably be raised regarding which publics and decision makers will be informed, and in what format. Shall the report be solely between the teacher and supervisor (and other administrative staff in the district)? Or, shall the report be publicized on a within-school basis and thus be available to the teacher's colleagues? Or, finally, shall the report on all teachers in the district be publicized more broadly, perhaps to the community at large? Accountability implies a public quality in the system, but the form in which public disclosure will take place is not explicit in the Stull Act. Public reporting is likely to have great influence on the quality of teachers who remain in the district. Aside from the departure of poorly qualified teachers who wish to avoid notoriety, it is not clear what other effects might accrue from public reporting of the results of the teacher evaluation process.

What is perfectly apparent from the preceding discussion if that the instructional improvement function and the teacher judging function, while relying on common evaluation information, are two distinct decision contexts.

Instructional Improvement and the Stull Act

In the previous section, the judging or evaluation function of the Stull Act was discussed. As has been noted, the same data serve another purpose--instructional improvement. Thus, while some of the same material on evaluation discussed by other speakers has been reiterated in this paper, it is necessary to understand the evaluation functions in order to understand the data available for initiating instructional improvement.

The measures that might be used to identify deficiencies in teacher performance should be examined again. The nature of the measures used in the evaluation has great impact upon the possibility of prescribing instructional remediation. Each type of measure carries with it not only the utility of its data for instructional design decisions, but also practical considerations related to the cost of using the measure.

The Question of Adequate Measures

There is no sense in detailing the conflicts in viewpoints regarding appropriate data sources for teacher evaluation. Those individuals who will concede that teaching is amenable to external evaluation are seriously dogmatic about the procedure they feel is best suited for gathering evaluation information. Thus each position will be slightly overstated. Some argue for measures of pupil growth against specifically stated microobjectives. Their belief is based on the assumption that teachers either know or ought to know how to teach these goals and that merely because of the articulation of the public's concern the teachers will redouble their efforts and start producing. Others assert that measures of teacher behavior derived from observation of daily activity represent a better approach to the problem. They justify their biases by using previously validated scales that are based on identified theoretical assumptions. Other equally ardent groups faithfully support the use of standardized achievement tests to provide classic, circumspect data regarding teachers' effects in the classroom. The variety of other teacher evaluation suggestions encompasses the use of collegial evaluation on poorly explicated criteria as well as the trial of laboratory-like performance test situations.

Before proceeding, one brief, tangential comment is necessary. Decisions regarding data source for teacher evaluation in past academic research have been academic and at best inconsequential (assuming that these are different). Researchers and teacher trainers adopted the indicator of teacher effectiveness that was likely to be favorable to their hypotheses or notions of program development. The consequences of error, though personally meaningful for the individual investigator, were negligible or nonexistent for the profession at large. Teacher evaluation has not been characterized by breakthroughs in concept or technology, but the selection of adequate teacher evaluation measures is especially critical at this time. If unworkable measures are selected, and the contribution of teachers in classrooms is subsequently discredited, the teacher accountability era may be the shortest-lived "innovation" of all. Hopefully, measures can be identified and applied which comply with the accountability needs of the public and are responsive to the tasks that teachers are best able to perform. Because of the anomalous situation in which teachers and their supervisors are presently placed, it seems that a sound precept for the implementation of teacher evaluation measures might be as follows: Because there is no single *measure that is likely to be useful, teacher evaluation must be based on multiple criteria*. The use of multiple criteria for teacher evaluation is recommended not only because it represents a safe procedure but also because it permits the evaluator to keep options open with respect to how the terms "deficiency" and "improvement" will ultimately be interpreted.

Types of Measures

In considering the alternatives for estimating teacher effectiveness, two major categories present themselves: (1) measures of pupil outcome and (2) measures of teacher behavior. In reflecting on the existing alternatives within each category, a natural distinction is apparent between those measures that sample highly *specific* or particularized aspects of either pupil outcome or teacher behavior and those measures of a more *general* nature.

	Pupil Outcome	Teacher Behaviors
General		
Specific		

A matrix generated from these categories provides a convenient labeling system for available measures of teacher effectiveness. Each category will be discussed first by providing examples of each cell and second by describing costs associated with using measures included in each of the four subcategories. Conditions under which various types of measures are particularly useful will also be suggested. Finally, the improvement options feasible for each data type will be compared.

General Measures of Pupil Outcome

In the category of general measures of pupil outcome, the outstanding example is the well-known achievement test. These tests are often called "standardized," referring to the means of calibration for their interpretation, but they may be more properly labeled construct-referenced tests. Students are examined on a range of topics that apparently fall "in" a category. For instance, in the area of reading comprehension, test items may be included which sample the students' ability to deal with inferential statements, main ideas, sequence, and point of view. Because the tests undergo empirical development and revision, the characteristics intended at the point of test design may be distinct from the characteristics that are actually functional. The implications of such differences for instructional improvement strategies will be discussed later. Goals statements could be written at a level of generality so that available standardized measures could represent acceptable indicators of their achievement. Acquisition costs to school districts would be minimal, since test development has been conducted, if not initially funded, by the private sector. Training costs would be negligible, since all teachers in all districts have had some contact with procedures for administering such achievement measures.

Specific Measures of Outcome

The Stull Act emphasizes the measurement of achievement of specific objectives. Such tests ideally may be distinguished from standardized achievement tests in that a given set of items purports to measure a particular, operationally promise in October to develop a test for high-school English literature by December ought to be ignored (or questioned searchingly about their supernatural contacts). While development of domain-referenced tests is theoretically possible in a wide variety of subject areas, the test developed is also limited by the level of analysis conducted by subject matter experts. As noted elsewhere (Baker, 1972), analysis of given content areas has not reached the level of sophistication necessary for the design of domain-referenced tests, where clearly defined subject matter attributes are essential to the entire concept. Thus, properly constructed domain-referenced tests are not available in many subject fields and would be relatively costly to develop.

In addition, there are other serious costs that may be associated with their use. If such tests are used to supply the diagnostic data required for instructional improvement, more test time will need to be scheduled. Additional testing requirements necessitate training of teachers, not only to suggest valid reasons for intrusions into the instructional setting but also to prepare teachers in methods to reduce the salience of the testing process itself and to enable them to conduct necessary assessments in an atmosphere of casual routine rather than hyperactive anxiety.

General Teacher Behaviors

A third type of measure appropriate for teacher evaluation derives from reports of general patterns of teacher behavior. Subjective ratings by principals or other colleagues provide one form of data that may be properly considered in accountability problems. Teachers may be rated in terms of their approximation to the ideal teacher, with what constitutes "ideal" left unspecified. Ratings of traits such as fairness, openness, classroom control, and relationship with pupils could be used as specifiers of the dimensions of "ideal."

Another kind of measure that falls into the general category is a report of the teacher's activity patterns: readiness for work, preparedness, range of materials he or she has available, and other indices of "hard work."

The advantages of measures in the general behavior category are few but important. For example, the cost of developing and employing such measures is low. Little training is required for such general reporting to be readily accomplished, especially if principals provide the ratings. On the other hand, since certain of the categories may be difficult to operationalize, any reported negative judgments will be hard to substantiate and may further complicate the problem of communicating identified deficiencies to the individuals involved. A problem, of course, is the effect of possible collegial collusion. Colleagues and administrators may be reluctant to criticize a co-worker, and inflated ratings may result. The alternative, to use "more independent" raters, is unfortunately not feasible, not only for cost reasons but also because the amount of time necessary to sensitize such raters to the "traits" of the evaluated teacher would not be available in real situations.

Specific Teacher Behavior Measures

Only a brief note of the category of specific teacher behaviors will be made, since the topic is amply treated in the presentation by Drs. Flanders and Morine. As in the other categories, there is a range of specificity of observations that might be made. On the one hand, one might use particular observation instruments developed to observe the pattern of teacher-student interaction. Another type of specific teacher behavior observation could center on the teacher's usage of a given method in the teaching of a subject matter area--for instance, the Moffett approach to the teaching of English in secondary schools. A more constrained form of the specific teacher behavior measure might be the observation of prescribed routines required for the use of validated programs. Thus, a teacher who followed the directions for implementing the SWRL kindergarten program would receive a high rating.

The costs of adopting an observation of specific teacher behavior as an evaluation data point are clear. First, there are training costs associated with employing or preparing well-trained observers to employ the scales of interest. In addition, there are some psychic costs to the observed teacher, because the data collection process is obtrusive and, unlike pupil outcome measures, the teacher cannot talk himself or herself out of responsibility for the results of the observation. Since the teacher's behavior is presumably under his or her own control (unlike outcome data), no other individual can be assigned culpability for poor performance. Because poor materials or recalcitrant or irregularly attending students cannot share the credit or burden of poor ratings, the bases for ratings are usually well delineated. The teacher cannot transfer any blame to the raters for their "misinterpretation" or "maliciousness," as is possible in the use of general teacher behavior ratings. So one might suggest that, of all possible data bases for teacher evaluation, the use of specific observation data is the procedure that places improvement responsibility most directly within the control of the teacher.

The four types of measures differ not only in intent and cost but obviously in the quality and utility of the data they supply. The entire cost of implementing a teacher accountability system should be considered; this includes the start-up and maintenance cost of collecting data, the analysis costs in terms of how easily the descriptions are translated into judgments, and the ultimate cost-effectiveness--how feasibly and with what probable success data from each type of measure can be translated into improvement interventions.

Improvement Strategies

Because the ultimate success of any accountability system resides in the ability to remedy identified deficiencies, the question of the relationship of evaluation measures to instructional improvement strategies is critical. How does one sort through the variety of data types that may be employed and reconcile any selection with the multitude of improvement enterprises which might be required? Certainly, there is no virtue in adopting any single approach to the improvement task. The problem of selection becomes particularly awesome after careful study of the literature in instructional research, for the paucity of recommendations for lines of clear action characterizes the summaries of most investigations.

In an attempt to simplify the decision process--for decisions will be made despite the proclaimed difficulty--it may be useful to conceive of instructional interventions in the following schemes. Available evidence indicates that there are instructional improvement strategies that are likely to succeed when compared with some whose effects have not been explored (or if so, without positive results). The first sorting results in a division of improvement strategies into categories of high probable effectiveness and effectiveness unknown. Instructional improvement strategies can also be separated according to the cost of employing the intervention. Cost estimates must consider not only the amount of money and time necessary to institute training of teachers and to acquire and implement any requisite student materials, but teacher time as well.

These two factors in considering strategies are shown in the matrix below:

		EFFECTIVENESS	
		HIGH PROBABLE	UNKNOWN
COST	LOW		
	HIGH		

To illustrate this design, some examples of each type of improvement strategy will be discussed briefly.

Low Cost--Effectiveness Unknown

This category of improvement strategy is frequently applied in the schools. For example, teachers in some districts are provided with a released-time allotment during which they obtain "in-service" training. This activity permits the teacher to visit teachers in other districts who are trying something new or to attend workshops or exhibits on topics of high interest. The cost of such activities is relatively low; the results in terms of instructional improvement are generally unknown. Effects of a mental health nature, however, are presumed to be high.

High Cost--Effectiveness Unknown

Less frequently found than the category just described, the high cost--indeterminate effectiveness intervention is likely to be seductive to people involved in instructional improvement, partly because of societal belief that things that cost heavily must be good. A priori, a Mercedes is better than a Toyota, although both come from technogically advanced societies. Purveyors of untried but innovative instructional approaches feed on similar prestige needs. If performance in natural science is low, does it not follow that providing teachers with training in applying new ideas and using expensive, albeit untested, classroom kits will be likely to improve the situation? Should not teachers who are having difficulty in teaching be exposed to year-long sequences of group process training? Too often the answer turns out to be unequivocally, unsubstantiatedly "yes." Perhaps "unknown effectiveness" should not be considered in a discussion of improvement strategies. But the present concern is with representing what is likely to happen in the name of instructional improvement, and suggesting the rightful place of each of the categories in the complex accountability situation.

High Cost--High Effectiveness

The critical dimension for all interventions that may be classified as high-probability effective is the existence of a body of evidence supporting the effects of their use. Some instructional improvement strategies can be implemented with the probability that desired behaviors or subsequent outcomes will clearly be remedied. Examples of this class of improvement are instructional programs that have undergone careful empirical revision during development. There is ample evidence that many of these programs work in a variety of settings. These are often referred to as validated programs. Programs that are designed to affect teacher behavior, such as those developed by the Far West Laboratory, represent one kind of high cost--high effectiveness artifact. The programs are high cost, not dramatically in terms of obtaining a program or even in requisite equipment acquisition, but in terms of the length of training time required for a teacher to modify even a relatively limited set of behaviors. However, there are data to support the programs' effectiveness. Another high cost--high effectiveness program would be one that is directed toward the children. For example, if a school district found that its teachers shared a deficiency in the effective teaching of beginning reading, one way to remedy this would be to acquire a validated program for use by the children in the schools. The cost of implementing this strategy would focus, not on training teachers as in the previous example, but rather on the capital costs of material acquisition for the learners themselves. In other cases, the teacher would become the program operator. Likelihood of success would be high if the program procedures were followed, and improvement would be an expensive although real attainment. The critical dimension for all interventions that may be classified as high-probability effective is the existence of a body of evidence supporting the effects of their use.

Low-Cost--High Effectiveness

One must suppress one's tendencies to disclaim knowledge of any such kind of instructional improvement strategy, for there are some known procedures that may be so characterized. Certain ideas in the field of instruction may, if one is generous, be elevated to the status of principles. Studies of such principles are plentiful and a growing research base attests to their utility across a wide range of tasks, subject types, and settings. Such instructional principles may be discerned in the programs of demonstrated effectiveness produced by well-funded instructional development agencies. These principles, derived from research in both operant and cognitive psychology, represent but a few of the principles that might form the substance of low-cost training. Examples are direct practice, where the learner is provided opportunity to practice criterion-type tasks; knowledge of results, where the adequacy of responses is communicated to the students; and task description, where the learner is informed of the nature of the goal. Under certain conditions, training in the use of given instructional principles may represent the best instructional improvement strategy available, for principles prescribe only the underlying structure of teaching; teachers are free to couch the principles in the presentation mode of instruction that is most comfortable for them.

While this intervention is categorized as "low" cost, low is obviously a relative term. Teacher training and practice sessions will need to be arranged. There will be slippage as each principle is rendered in a unique way by teachers. More importantly, incentives for the teachers to adopt the use of the desired principles must be presented in concrete forms. Those responsible for teacher evaluation and improvement must find methods to reward teachers for altering instructional dispositions and actions in desired directions, and the use of such principles is likely to pay off under certain evaluation conditions.

How does one negotiate between the two matrices suggested, selecting data likely to be amenable to an instructional improvement strategy? For each of the various data types, what are the improvement strategies available, and which of them is likely to be capable of actually improving things? The success of any of the four varieties of improvement strategies is inextricably linked with the prior decisions regarding the type of measure employed in teacher evaluation. These relationships are further modified by the particular context in which teacher evaluation takes place. Let us consider some expedients.

Outcome Measures and Instructional Improvement

If, for a particular subject area, performance on a general measure of pupil outcome is found to be deficient, what are the would-be improver's alternatives? Because information on such measures is presented in terms of the individual's or group's overall status with regard to a topic, those responsible for improvement will find it difficult to focus on particular skills in need of improvement. The only alternative will be a dramatic change in the overall method by which the subject is taught. Obviously, any method or procedure, such as a materials-based curriculum, which has data to support its effectiveness would be a desired choice. Unfortunately, the choice is expensive as well. Any other choice of improvement strategy has very little chance of payoff for the measure employed. For instance, training teachers in the application of such low-cost alternatives as the use of practice and task analysis skills presupposes that the teacher has specific information regarding what is to be practiced or what skills require further analysis. General measures of pupil outcome do not provide the necessary framework for the successful application of such skills. The option is still low-cost, but has an indeterminate payoff.

On the other hand, if an evaluator has employed a specific measure of pupil outcome, such as a domain-referenced test, the options for improvement strategies increase. Certainly, a total and expensive change in program could always be effected, but other, less costly interventions could conceivably yield the desired results. Because information on pupil achievement is presented in terms of subskills achieved and not achieved at desired levels, the use of validated programs might be limited only to those particular areas of deficiencies. For instance, instead of adopting a total program representing a completely new approach to reading, perhaps the use of short-term materials directed at the deficiency indicated by the data would be more appropriate. Thus, even for the high-cost, high probable effectiveness improvement strategy the use of domain-referenced tests could effect a saving. In addition, because of the careful design requirements of such measures, teachers could be trained to generate items analogous to those used on the domain-referenced instrument itself. Providing practice in similar, but not identical, items, would represent a low-cost, high probable effectiveness strategy. Of course, one must not overlook the possibility that performance levels might be poor on all subskills measured by a domain-referenced test. The choice in that eventuality is clear: either make an expensive and radical change in program or select particular aspects of the performance efficiency on which to focus the improvement (a lower cost strategy).

The intended message with regard to outcome measurement is summarized as follows: costs are associated with both ends of the accountability decision--in the selection of measures and in the implementation of improvement strategies. Costs may be cut by using a more accessible, more familiar sort of outcome measure. It is expected, however, that such a decision will necessitate the selection of a high-cost improvement strategy. On the other hand, a domain-referenced measure, although costly to acquire, may permit the employment of high-effectiveness, low-cost, improvement techniques.

Teacher Behavior Measures and Instructional Improvement

Relating teacher behavior measures to classes of instructional improvements is at once a simpler and a more difficult task than the consideration of the pupil outcome area. It is simpler because, if general measures of teacher behavior are selected, almost anything could be classified as "assistance" in improving performance and almost nothing will be likely to work. It is more difficult because cost factors for instructional improvement strategies do not favor the specific measures as they do in the outcome category. Costs for improvement strategies may range widely, depending upon the particular aspect of teacher behavior that has been observed.

A common way to select measures is of course to identify aspects of teacher behavior felt to be critical by teachers, experts, and administrators, to observe the teachers against these criteria, and to use whatever happens to be available (if anything) to provide improvement assistance when the time comes. A very different strategy may be suggested for selecting the actual dimensions of teacher behavior to be observed. One might use a backward chaining method, adopted from early programmed instruction lore, and scan available, successful teacher-improvement programs which fall within cost constraints. These options prescribe the dimensions of teacher behavior that may be selected for observation. The benefit of such a strategy is to increase the probability that the improvement cycle requisite for "deficient" teachers will result in ultimate success.

Types of Measures, Instructional Improvement, and the Real World

At an earlier point in the paper, the position was taken that multiple criteria ought to be employed in teacher evaluation and that there was merit in each of the four data types. The subsequent discussion of instructional improvement strategies should have raised some doubts about the merits of various data types, for probable improvement cycles depend heavily on specific data. Perhaps we have only a row instead of a matrix.

If one will grant, for a moment, that the best sort of data to have is specific pupil outcome data, then why fool around with anything else? The answer resides in a parable.

> Once, in a far-out land there lived the Good Evaluator. This Good Evaluator visited the schools to help out as necessary. The Good Evaluator wanted to be a change agent and through magic was able to develop and refine domain-referenced achievement tests in each area of instruction. With an extra dose of magic, the Evaluator was able to schedule tests as often as they were required, and was even able to have the children like it. When the data from the domain-referenced tests were analyzed, however, the Good Evaluator was sad, for many teachers were deficient in many areas of instruction. Improvement programs were necessary in each area of instruction. The Evaluator's magic was all gone.

And that's the end of the story. Even with incantations, talismans, and sleight-of-hand, there is no way for districts to mount serious improvement programs in all areas that are likely to show up deficient on domain-referenced achievement tests.

The alternative is to invoke the notion of goal accountability and to select areas of instruction that are of high priority. Suppose reading and mathematics were subject areas selected for intensive concern in the elementary schools. Domain-referenced achievement tests would be provided in those areas only. Improvement programs of high probable effectiveness would be selected, either a high-cost renovation of existing programs by adopting well-tested materials-based curricula, or training of teachers in the use of principles which, when applied in conjunction with specific performance data, are likely to result in improved performance. For other areas of instruction, general measures of pupil outcome could be employed. Deficiencies might be operationalized at less stringent levels in these areas. These improvement programs could be of the low-cost, indeterminate-payoff variety. In a situation where priorities are set, the choice of low-payoff improvement strategies is not only acceptable but certainly recommended. One would expect that, over time, priorities would be reevaluated as certain subject areas became less problematic for teachers.

Another situation in which domain-referenced achievement testing might not be most useful is the following. Suppose, in certain urban secondary school settings, daily pupil attendance is below 50 percent, and not always the same 50 percent. Pupil outcome measures are less meaningful since students may not have had sufficient opportunity to interact with instruction. (Purists might say that the school should make itself a more reinforcing place, and one would agree, but that is often a separate problem from teacher evaluation.) In order to comply with the law, a low-cost pupil outcome measure should be selected, for there is little reason to administer costly diagnostic instruments when students will not be present consistently enough to profit from remedial instruction. This is a hard line to take, particularly as it makes explicit the notion that teacher improvement is not the only problem in such settings. Data for accountability purposes should focus on teacher behavior indices, and improvement programs should concentrate on training teachers to exhibit the desired behaviors.

In summary, the discussion regarding the relationship of measures and instructional improvement strategies should have identified the costs at the points of selection and use of the measures as well as the costs of improvement strategies. Evaluators will have to decide at which point they wish to effect savings. The recommendation of this paper would be, in general, to identify high-priority instructional areas and to use domain-referenced tests in those areas. Domain-referenced tests are suggested because they may permit the analysis of "deficiency" at a more refined level and perhaps one less expensive to service. General outcome measures tell you whether your school is good or bad; if it is bad, there are few options but to renovate the entire program. On the other hand, domain-referenced tests are not exclusively recommended, even if tests in a variety of areas were available. Too much confusion would follow the identification of multiple deficiencies in multiple areas, resulting in a situation with which the district or school could not cope. If, in addition, there is a particular desire to focus on teacher responsibility per se, rather than on improvement of school performance in general, then the use of teacher behaviors is recommended, for teachers, to the extent that we all do, have their own behavior under their control. This paper recommends selecting a few high-priority areas and gathering domain-referenced data in these. For the less critical areas of instruction, easily available and administered standardized achievement tests will effect compliance with the law.

Inadequacies of the Stull Act

Suggestions, recommendations, second guesses, and discourse have been focused on the implementation of a particular law in this state. The time to discuss the adequacy of the law is at hand, and instead of discussing how to survive under its present mandate, the paper will address deficiencies the Stull Act has to remediate. The inadequacies in the Act include errors of both omission and commission. For instance, the timing of the Stull Act seems to be somewhat inappropriate. The sequence of progress in implementing the various accountability systems in California seems to be following a rather strange order. That is, the Stull Act, which is concerned primarily with outcome accountability, is moving ahead very rapidly in this coming year; in contrast, goals determination (goal accountability), as exemplified by the Statewide Joint Committee on Educational Goals and Evaluation, has floundered for several years and now the ball seems to be passing to local school districts. Presumably there will be some attention to this area in the next year, but in all likelihood it will not be nearly as "pressing" as the Stull Act. As a result, in most school districts attention to goal determination will follow outcome evaluation. Meanwhile, program accountability, as exemplified by program validation on a systematic basis, sits almost unnoticed.

To say, as is said in the Stull Act, "given a teacher who does not measure up, tell him or her what to do in order to improve," is in essence to place outcome accountability before program accountability. One may hypothesize that an instructional program (the materials, products, and procedures used by a given teacher in combination with the individual efforts of that teacher) was not satisfactory in producing the desired educational outcome and that the supervisor or evaluator of the teacher is asked to make recommendations as to how to rectify the situation. This is preposterous! If the principal did not know what program to use (what validated programs existed) at the beginning of the year, what makes anyone think that he will be in a better position to make prescriptive statements at the end of the year?

Another area of potential ambiguity in the Stull Act deals with the responsibility of the teacher in the period following that in which recommendations are made. Suppose, for example, that a teacher has performed poorly relative to the stipulated outcomes and that, in terms of the law, "(the) employee is not performing his duties in a satisfactory manner according to the standards prescribed by the governing board." The supervisor then confers with the employee (teacher), making specific recommendations as to areas for improvement in the employee's performance. The employee then follows the recommendations established by his or her supervisor during the coming year and, by any process measure used, it is determined that the teacher did what the principal said he or she should do in order to improve. What is to happen if at the end of the second year the evaluation of the teacher on student outcomes is no better? Does the responsibility for poor performance rest upon the teacher or upon the principal who may have given an inadequate prescription? This is a difficult question, and it seems to be one that has been largely ignored in the discussion of the Stull Act. This issue relates to the general consideration of distinctions between program accountability and outcome accountability discussed earlier. If there were validated programs (for which evidence existed about the anticipated outcomes under specified conditions), then the prescriptions of such programs should yield the specified outcomes. And if the prescription does not yield such outcomes, it clearly is the fault of the teacher, who must either be operating the program incorrectly or simply be incapable of using the program. This would be a relatively clear-cut case, for the program has been validated and has demonstrated that it can accomplish the specified outcomes. However, such a case would be rare today because of the lack of available validated programs. Thus, when a principal suggests a prescriptive activity to teachers, with no validated programs available, where does the fault for nonattainment of student outcomes lie? The criteria will ultimately rest, one suspects, on some base such as "general agreement on the reasonableness of the prescription"; that is, most professional educators would concur that the prescription recommended or the assistance provided was reasonable, given the situation.

What is clear from all of the foregoing is that remedies or prescriptions will not be available for all situations. One of the constraints of instructional programs that depend upon people is that each person has certain characteristics and abilities that are not modifiable. An analogy can be made in terms of the famous television character who, no matter how hard he tries, continues to get the response, "Sorry Charlie, you're not good enough for Starkist." One anticipates that the Stull Act will identify a few Charlies, but only a few.

Another serious problem with the Stull Act is the manner in which it is likely to be implemented. Although our society has been characterized as one of rapid technological change, one cannot infer that we should arbitrarily attempt to race forward with every possible improvement. Analysis of side effects is a common part of most technologies, and the present concern with pollution makes us all wish someone had thought of it sooner. The side-effects analysis of the Stull Act is incomplete at best. If the accountability system could have been implemented developmentally, certain negative consequences which are imminent might have been avoided. If the Stull Act fails to accomplish its ends, it may do so not because the idea was wrong, but because of a simple procedure that was misunderstood, or for some other inane reason. A leisurely pace is not recommended, but careful trial and revision, in other words, development, is necessary for anything (except some lyric poetry) to work.

Summary

This paper attempted to deal with the notion of accountability and instructional improvement. Accountability was defined and separated into three parts: goal, program, and outcome accountability. The relationship of the Stull Act to other accountability activities was explored. The evaluation requirements for accountability were briefly described. Four types of measures suitable for holding teachers accountable under the provisions of the Stull Act were outlined: (1) general measures of pupil outcomes; (2) specific measures of pupil outcome; (3) general measures of teacher behavior; (4) specific measures of teacher behavior. Each type of measure was briefly described and discussed in terms of costs, both salient and hidden. General strategies for instructional improvement were also described and classified into four categories by probable effectiveness and level of implementation cost. The relationship between data type and improvement strategy was considered, and available alternatives for each measure were briefly discussed. These concerns were overlaid against some of the constraints which are likely to operate in the schools as the drive to evaluate and improve accelerates. Recommendations were made with respect to the selection of data types and improvement strategies. Finally, the Stull Act itself was examined, and questions regarding implementation were raised.

The paper attempted to convey in the treatment of all of these topics a concern for the instructional improvement of the schools. The need for evaluation should not hopelessly entangle and alienate administrators and teachers; it should not worry students. It will take some time to see if the idea has merit. If it fails in California through careless implementation, no one will get a chance to test the concept again for a long, long time.

References

Alkin, M. C. Evaluation and decision making. Planning and Changing, 1973 (in press).

Baker, E. L. Using measurement to improve instruction. Paper presented to a symposium, "The Relative Strengths of Norm-Referenced and Criterion-Referenced Achievement Tests," at the annual meeting of the American Psychological Association, Honolulu, September 1972.

Hively, W., Maxwell, G., Rabehl, G., Sension, D., and Lundin, S. Domain-referenced curriculum evaluation: A technical handbook and a case study of the MINNEMAST Project. CSE Monograph Series on Evaluation, No. 1. Los Angeles: Center for the Study of Evaluation, 1973 (in press).

Husek, T. R. and Sirotnik, K. Matrix sampling. Evaluation Comment, 1968, 1 (3).

Klein, S. P. and Alkin, M. C. Evaluating teachers for outcome accountability. Evaluation Comment, 1972, 3 (30), 5-11.

DISCUSSION OF PAPER BY ALKIN AND BAKER

THOMAS W. BUTTERWORTH
Los Angeles County Schools Office

I feel I am placing myself in a rather peculiar position regarding Dr. Alkin's and Dr. Baker's paper. The standard role of a discussant is to critique--to make constructive comments in a very specific manner. I find the paper so exceptionally good that it is hard to discuss it critically. There are a few points here and there, but they are more a matter of emphasis, I think, than of substance. I will comment on a few of these points of emphasis in a few moments.

I think the paper comes at a very propitious time in the conference. Initially, perhaps, we were overwhelmed with highly sophisticated statistical explanations of error rates and bias in terms of using various kinds of instrumentation in measuring pupil behavior. But I think that the Alkin and Baker paper presents a very neat, very comprehensive, and very well-integrated total system of evaluation, or total system of accountability.

I would like to digress for a moment and examine some of the impressions I gained from the earlier discussions. I believe these impressions underscore the need for the kind of total accountability proposed by Drs. Alkin and Baker. Now it is obviously a tricky business to suggest that my subjective impressions of what earlier presenters and discussants said, either explicitly or implicitly, constitute facts. I hope I don't do violence to any of the presenters or discussants. Obviously, I have made a rather random selection of some of the points that were made. With that caveat, I would like to mention some of these impressions, not necessarily in order of importance.

One of the things I thought I heard being said is that evaluation can occur in the absence of data. I suspect that if one is a metaphysician or a psychic, maybe this could work. But it would be hard for me to imagine evaluation occurring in the absence of any kind of instrumentation, unobtrusive or obtrusive, formal or informal. Somebody has to view or extract some data before he makes some judgment about its meaning. He may not be aware of it but he is doing it. Another impression I got was that the research, statistical, and instrumentation problems are so great in attempting to relate teacher behavior to pupil effects that we give up the quest and dispense with pupil outcome evaluation. Instead, we determine instructional effects by gathering data on teachers rather than on pupils. Another impression gained is that there is much more to a teacher's job than changing children's behavior. If this is true, then I don't think we can say that what the teacher is doing is teaching. (A colleague of mine refers to this as taking money under false pretenses.) Another impression is that since we cannot use child behavior as a criterion for instructional effects, we should change the purpose of the schools to meeting the needs of the adults who work in that system. At least we can make them comfortable, even though they are not sure what the students are learning.

Another impression is that we might have the children judge teacher effectiveness in terms of such traits as human decency. This seems to me to be a less valid and less reliable measure than having principals make such judgments, as they have usually done. It is almost like a client telling the psychologist his impression of the psychologists's strategies as a criterion of how effective he has been with the client.

Also, I gathered that some people feel the schools do not have a significant impact in reaching the objective they would seem best qualified to achieve, namely, cognitive development.

Therefore, the inference is, let's at least make the place pleasant enough so that students can feel good about themselves. (Another colleague has indicated that sometimes this ends up by teaching children how to distort reality. That is, even though they are cognitively incompetent, at least they feel good about themselves.)

Another impression I gathered is that pre-specifying outcomes in terms of pupil behavior will lead to a sort of brave new world. To counteract this it is suggested that we should have a "surprise curriculum," not because we plan for surprises, but because they will randomly occur. Also, I gathered that accountability shouldn't receive the attention it is getting because it implies a lack of trust in the natural healthy development of children. It was also inferred that accountability will lead to over-control, focusing on trivial outcomes. Finally, I learned that our clientele should trust us to do the right thing (whether we are talking about the client as the child or the parents).

The reason I mentioned these impressions in a rather dramatic (and at times facetious) manner was to indicate what I feel is a confusion about what accountability is. I think this confusion can be rather effectively resolved by means of adopting the suggestions in the excellent paper by Drs. Alkin and Baker. Of course, there are varying interpretations of what the real problem is. Is the problem one of dealing with the Stull Act? Certainly, since it is law and it must be confronted. But if we see the Stull Act as the beginning and the end, we do not grasp the essence of the real problem. I see the Stull Act as only one part of a mosaic of strategies being employed in attempting to deal with a publicly perceived, growing ineffectiveness of schools. PPBS, vouchers, performance contracting, the Vasconcellos goal setting activity--these are all part of the same movement. A directly related spinoff, of course, is the growing lack of credibility the schools have in the public eye.

We have always talked about doing systematic evaluation. Now we are being forced to do it when in fact we should have done it prior to this time. Leon Lessinger has many times repeated the phrase, "It is the need to produce results, not promises." The fundamental problem, then, is how can we improve the effectiveness of what we purport to do, which is to teach or to change learner behavior? And how can we report this in a meaningful manner to our clients--the public. As Dr. Popham pointed out, we are just beginning to do the kind of research and development that will permit us to be held accountable, and in the process, to improve our strategies of instruction. And yet, what I heard yesterday was not that we should continue to build on what Popham and others are going, but that we should beat a hasty retreat to the former practice of asking public indulgence and trusting that we will do what is right instead of rightfully producing results.

The paper presented by Drs. Alkin and Baker addresses itself to this problem of meeting the commitment we should have to our clientele by designing what I would like to call a "total performance system" which integrates all the essential elements of evaluation. At the outset, it acknowledges that the Stull Act demands an immediate solution, but it emphasizes that the solution we create must address itself to the broader context of accountability. The solution is proposed in such a way that fairness to certificated personnel must be considered in making judgments about their effectiveness. Constraints both within and outside the system beyond the control of the participants are identified.

I was happy to hear Dr. Flanders also endorse a more systematic type of planning, and of course the flavor of systems analysis runs throughout the paper by Drs. Alkin and Baker. Constant attention must be given to the public and school interface through goal accountability, admission statements, recycling to ensure relevance of strategies to target objectives, and constant examination of alternative strategies, with necessary tradeoffs made in cost-benefit terms. Consideration must also be given to varieties of multiple consequences. All of these things form part of a well-thought-out proposal.

I would like to refer to a couple of points in the paper that are merely, perhaps, a matter of emphasis. Using the term accountability to refer to disparate means-ends relationships in a management system bothers me. I acknowledge that there are many meanings attached to the word, and I also acknowledge that Drs. Alkin and Baker have, I think, a different meaning in terms of the kind of data produced for those three different types of accountability--outcome, program, and goal accountability. I see accountability as referring to the ability of the school to publicly verify the results it is producing, and of course to improve the effectiveness of its programs in changing learner behavior. In this sense there can only be outcome accountability, for outcomes in terms of learner behavior are the primary target. All other functions constitute means to that end. In practice, I find people are not making this distinction.

Organizations are wildly writing objectives for all persons in the system. Dr. Popham has rather jokingly called this "an objective circus." Very few people are giving much consideration to the question I am fond of asking, "What is an objective supposed to do?" Not, "What is an objective?" Anybody can specify the components of an objective. But when we talk about the kinds of decisions we need to make, then to answer the question in terms of what is the criterion we use to judge if we have a performance objective or if we have some other kind is, I think, critical. People are treating a number of decision-making components as though they were performance objectives, which would mean that they should produce information about the effects on outcome behavior and thus judgments of program effectiveness. Rather than being performance objectives, they are actually standards for judging the merit of strategies or of what people are doing in the system. In the absence of performance objectives this prevents the testing of alternative strategies against performance outcomes.

A brief word is in order here about the concept of the validated program. It's true that much benefit is derived by "pretesting" or validating a curriculum sequence in another setting. However, the problem of bias remains in making statistical predictions of learner improvement with a given sample.

I was very impressed with the treatment in the paper on staff development strategies in terms of relative cost and payoff and the need for setting priorities in terms of goal accountability.

As a concluding comment, I would like to warn us all that schools are in danger of becoming as irrelevant as psychiatry. Richard Stuart of the University of Michigan, in a clever and scholarly book called Trick or Treatment, discusses a concept usually applied to medicine. It is called "iatrogenic." Translated, it means treatment-produced illness. There is a parallel between the illness apparently produced by psychiatry and the failure apparently produced by the schools. In both cases, research suggests that if you are a group of clients, or students, on the average you will fare about the same by exposing yourself to the treatment or by not exposing yourself to the treatment. However, if you are a given client or student, you will get either much better or very much worse after being exposed to the treatment. In other words, the rich get richer and the poor get poorer. Jencks is right, as Dr. Levin said. The schools are taking poverty children and making poverty adults of them. The tragedy of this is that we are assuming that there is little more that can be done.

Alkin and Baker's paper, it seems to me, provides the framework within which alternative strategies can be monitored and tested against the criteria of outcome accountability. It is only in this way that we can begin to see where we are going, the things that we are doing right, and the things we are doing wrong, and thus make the school more than just a random (or perhaps a nonsignificant) effect on human development. Stull has come and Stull may go, but the problem which preceded the Act will remain. Accountability is not going to go away.

BERNARD H. McKENNA
National Education Association

Professors Alkin and Baker have done well, I think, in describing and illustrating the relationship of teacher accountability to the other accountability levels--that is, to what they call school and instructional constraints. They have also introduced what appears to be a new term in this conference: goal accountability. They would make the school board accountable for selecting appropriate goals and for assuring the public that these goals and their accompanying objectives are being pursued in the educational program. But they don't tell us quite how the school board will be judged on this accomplishment. Presumably they will be judged by the public in some way. But should not educators and students also have a role in studying goals and selecting objectives? And if so, isn't goal accountability a broader responsibility than Professors Alkin and Baker suggest?

Needs assessment is an important part of goal and objective setting, I would assume. In many places, needs assessment is very much a function of the entire school staff and of the students. So should they not share in the responsibility for determining appropriate goals and objectives? What I am suggesting is that goal accountability is indeed a broader responsibility than was indicated in the paper by Professors Alkin and Baker.

I had hoped that, under the heading "Evaluation to Improve Performance," we might have learned whether performance evaluation is different from evaluation to determine change of status. (By change of status I mean retention, promotion, and tenure.) Professors Alkin and Baker say that these kinds of evaluation are indeed different, but they don't pursue that point very much. I also hoped we might learn whether evaluation to improve performance and evaluation for retention and promotion are highly compatible, whether they are mutually exclusive, or whether they are only loosely related. The Stull Act, after all, mandates assistance to upgrade performance for those personnel whose competency is not up to standard. So it would appear that the main purpose is to improve performance.

We have a problem here in not distinguishing among several different kinds of things: (1) evaluation for change of status (retention, promotion, tenure); (2) evaluation for improving instruction (this could be very broad; it could include program evaluation); (3) evaluation for improving the performance of certificated personnel; and (4) evaluation for improving student learning. Obviously we would expect there to be a strong relationship, but since the relationships are not very clear some independent selection among them seems to be defensible.

The language of the Act appears to allow considerable leeway for choosing among evaluation for change of status, for improving instruction, for improving performance, and for improving student learning. I say this because the Act speaks of assessing personnel competence as it relates to the standards of student progress. "Relates" is a general enough term to provide for a wide range of interpretations that could lead to constructive action. This is what Dr. Glass was getting at when he said there will be strict-constructionist interpretations of the Act and there will be liberal interpretations.

Mr. Stull has said that local school people need to get together to work out the implementation of the Act, and that it isn't highly prescriptive. So you Californians may find it more manageable than would appear at first examination. I don't mean that you shouldn't ask the difficult questions concerning the relationship of personnel performance to educational outcomes. This only means you have leeway to begin implementing those things that are manageable while working concurrently to research and develop the others.

Professors Alkin and Baker say, and this is a quote, "Hopefully, measures can be identified and applied which comply with the accountability needs of the public and are responsive to the tasks which teachers are best able to perform." I am not very hopeful about this, and this is the crux of the matter. As I interpret what the public is expressing as its accountability needs, they relate mainly to raising achievement test scores. In many places they are even narrower than that--they relate to raising achievement test scores in mathematics and reading. I suspect that a high percentage of the California public, and for that matter the public all across the country, feel much as Mr. Stull does about this. I wasn't very hopeful when I came here that in the foreseeable future we can relate teachers' acts to students' achievements, and having heard Dr. Glass's fine analysis of the research, I'm even less hopeful than when I came. That being the case, it appears that for the present implementation of the Act, heavy reliance will need to be placed on criteria other than student achievement, at least as it is typically measured today by standardized tests. Let me repeat that this doesn't need to make one totally pessimistic about working toward relating teacher performance to achievement, especially under carefully controlled conditions where teachers are fully involved in developing the teaching strategies and the means for implementing them and for measuring the outcomes with students. The domain-referenced tests Professors Alkin and Baker cited appear to have some promise for this purpose.

The Alkin-Baker paper states that because there is no single measure which is likely to be universally useful, teacher evaluation must be based on multiple criteria; I strongly support this position. But I would add a dimension to the criteria Drs. Alkin and Baker use. Their criteria are input, process, and product (although they call them something else). I suggest that we must look at this along two dimensions--what is to be measured, and who is to do it. On the what side, it is pretty clear from the Alkin-Baker paper that we need to evaluate programs, we need to evaluate performance, and we need to evaluate outcomes. It says that all three are very complicatedly interrelated, and I agree. On the who-is-to-do-it side, we need to work on self-evaluation, peer evaluation, student evaluation of staff, parent evaluation, independent outside evaluation, and perhaps evaluation by other groups. It has been suggested that there are pressure groups in this country who evaluate schools all the time--the American Legion, the Chamber of Commerce, and so on--and we probably ought to get them involved as groups.

When we consider that we need to measure inputs, programs, and outputs and that we need to get all these groups involved, it is not only complicated but, as has been suggested here several times before, it is terribly expensive. Some papers have suggested that some kinds of evaluation are prohibitively expensive. I remember what Henry Dyer of Educational Testing Service said when that group was asked to take on the task of developing accountability tools when the New York teachers contractually agreed to demonstrate that they were accountable. In a New York Times interview, when asked "Can this be done?" Dyer said, "Yes--but it would probably take more time and cost more money than getting to the moon." If this is the case, we will have to believe that it is an integral part of instruction and so important that massive resources are allocated to it. I suppose this is not too helpful to those of you who must return to the firing line and live with present financial situations.

There is an analogy that is useful to me in understanding where we are and where we have to go on evaluation. I began teaching in a little town in the Middle West, and somehow I felt I needed an annual physical examination in order to be a teacher. It wasn't required by the school board, but I went to the only physician in town, and he did just four things. He took my blood pressure, listened to my heart with his stethoscope, put me on a table and pounded my knee to see if I had reflexes, and said, "Young man, you are fine for another year." Just recently I completed my annual physical examination in Washington, D. C., and it took three trips. The first time the doctor sat me down and asked what my relationships were with my family this past year, how I liked my work, and how I was feeling about myself--you know, a kind of minipsychological inventory. Then he took me in the examining room and did those three manual tests the old physician had done, but also about a dozen more. He put lights in my ears and chains down my throat with lights on them and a lot of other things. Then he said, "The second time you don't need to see me. We will make an appointment with the laboratory." I counted 14 laboratory tests, including electrocardiogram, cholesterol count, chest X-rays, several blood samples, and so on. The third time I went, all that had been graphed. Not only did the doctor have it graphed, he also had last year's test results. He said, "Your cholesterol count this year is thus and so; last year it was thus and so." My examination cost me $8 in the little Midwest town away back when; it cost me $150 this year.

I submit that in evaluation we are now about where that old physician was. If we really believe in it, it will require substantial resources, including large sums of money. And that's not much consolation to those of you who face fiscal crises almost daily back home. Maybe the most you can do is to work toward convincing your publics that their priorities are poorly ordered.

I own a vacation house on a lake. When I go along the road to the nearest village I see two snowmobiles, not just one, in almost every farmyard. But these people have voted down tax bond issues for schools more than once recently. I can't help but conclude that they would rather have snowmobiles than more teachers to individualize instruction. As long as schools are so ranked among the priorities of the American people there will be great difficulty in making the Stull Act work in terms of what I judge are minimal resources needed to do the job.

Alkin and Baker's paper gives some attention to what they call improvement strategies, and we need much more of that. In our panic about the overwhelming task of measuring competence, we sometimes almost totally neglect the question of what do we do after we have measured it. The most we have done, generally, is to counsel with the teacher (or the evaluatee) or to give him some kind of pencil-and-paper feedback. That is mandated in the Stull Act, but the mandate that goes beyond that, which requires providing assistance to upgrade performance, may be the most important part of the whole thing. If we assume that evaluation should be more for improving performance than for change of status, then the answer to the question, "What do you do after you have measured?" could lead to the necessity for mounting massive in-service education programs, the cost of which would probably make evaluation programs look very inexpensive.

Finally, I think there are six points that need to be considered very carefully in implementing the Stull Act. First, use multiple criteria, if at all possible. Even if it is only two instead of one, start with multiple criteria and don't pin all your hopes on one kind of evaluation. I believe that conforms with what Drs. Alkin and Baker suggested. Second, involve a wide variety of persons in the process. Third, don't expect much from the relationship between staff performance and student achievement for the present, but fourth, don't give up--work on it concurrently while implementing program and performance evaluation. Fifth, expect it to be expensive, and finally (I can't resist this, working where I do), don't just seek the advice of the certificated instructional personnel as it says in the Stull Act, but involve them fully from beginning to end, from goal setting to application.

EVALUATING EVALUATION SYSTEMS DESIGNED TO IMPLEMENT THE STULL ACT

GARFORD G. GORDON
California Teachers Association

A major question has arisen in the last 20 years about the effectiveness of education. For at least a century it was held that education was the key to a more satisfactory way of life. When it was observed that despite great increases in educational endeavors, the poor, criminal, and disaffected are still with us, the institutional purveyors of education were challenged. They responded by citing several defenses. Most commonly they cited the niggardly support given by society. Less often they called attention to the fact that institutional education operates over a very minor part of the life span of children and youth.

While the actual increase in resources committed to education was modest in terms of the total enterprise, the dollar amount was large. Moreover, it was raised primarily by forms of taxation which were the most obvious, inequitable, and onerous of all methods employed to milk the public for the financial support of social activities. There has, therefore, arisen a movement compounded of fiscal and social frustration, known as "accountability in education."

Those who support the movement have mixed motives. Most prominent among those not directly involved in education or work with youth is a desire to have the training requirements of future generations met without commitment of additional resources to schooling. Most prominent among educators is a desire to do a better job with those affected by educational processes. The former feel that a tight rein on the practitioners of education is needed; the latter feel that retraining and motivation for innovations are needed. Both believe that a careful study of what is going on and of how successful it is in doing what education is supposed to do is essential.

The Stull Act, the Commission on Educational Management and Evaluation, and the Joint Committee on Goals and Objectives represent attempts by California legislators to meet these concerns. The last is designed to determine what schools are really supposed to do. The Commission on Management and Evaluation is mainly intended to survey the mechanisms by which institutional education tries to do its job. The Stull Act is an attempt to tie personnel administration into these other efforts.

It is true that the various agencies and programs are autonomous, with the result that there is overlapping, inconsistency, and conflict. However, whether approached via the Stull Act, commissions on goals, objectives, and evaluation, or various forms of Planning, Programming, Budgeting Systems (PPBS), the purpose of evaluation is to examine what the relationship between resources committed and results obtained.

An immediate and superficial reaction to the Stull Act is to consider it as requiring only an assessment of the performance of individuals charged with responsibility for conducting the State's educational program. However, when the purposes of the accountability movement and the nature of the educational process are considered, it becomes obvious that the Act can only be implemented by establishing a comprehensive system for evaluating the entire educational program.

Any consideration of the implementation of the Stull Act and of the success of such implementation and the problems associated with it must be based on the recognition that evaluation of personnel performance is only one aspect of evaluation of the educational system. It is inseparable from evaluation of all other elements of the system. In order to judge the worth of attempts to implement the Act, it is necessary to apply some general principles of evaluation.

First, let us consider the nature of evaluation as applied to educational activities, using an approach as free as possible from adherence to any preconceived system or special terminology. We can then say that an evaluation is a judgment applied to a situation or to a process. It is often thought of as being applied to an individual. This application, however, is a secondary effect, reflecting the evaluation of the process or situation for which the individual is responsible. It involves additional considerations, such as the degree to which the individual has responsibility for or is able to influence the process or situation.

While evaluations can be made in many ways, valid ones require that the judgments be made on as accurate and complete data as possible about factors that have meaningful relationships to the situations and processes being evaluated. The meaningfulness of the relationships of any factor to a situation or process depends upon the objectives of the process or the aspects of the situation which are considered significant.

In evaluating educational matters, it must be recognized that there are two aspects of the function of educational systems. One is the process of education, which is designed to achieve certain objectives in modifying the behavior (in the broadest sense) of the students. The other is the maintenance of certain levels of physical and psychological well-being on the part of students, parents, and community during the time students are subject to the process. It appears, therefore, that valid educational evaluation requires consideration of both situational and process factors; and that judgmental standards should relate both to the objective of the educational process and to the environment in which it is conducted.

To apply these principles to educational evaluation it is first necessary to consider the purposes of schools. Of course, the specific purposes of a school system and of individual units within it--classes, schools, departments--depend upon the society within which it operates. However, certain things are common to all. Education is behavior modification --in other words, those who have completed their schooling or any portion of it should behave differently from the way they would have behaved if they had never been to school. Such modification of behavior may be trivial or profound; it may reflect conditioning or deep internal changes of perception, philosophy, and emotions. However, only the overt behavior changes that result will show whether or not an observable change has been wrought in the student. Anybody is at liberty to hypothesize internal changes not reflected in outward modifications of behavior but in the absence of outward manifestations, he has no way to show that change has in fact occurred.

A second purpose of the schools is the assumption of responsibility for children and youth over a significant part of their life spans. While it is hoped that activities conducted by the schools during this period will contribute to desired behavior modification, there are other objectives, including maintenance of social and physical conditions the community considers desirable for its younger members. Whether or not meeting the objectives in this area will promote--or even hinder--the attainment of objectives in behavior modification, this function of the schools is an accepted part of their role in modern society.

Any evaluation of a school system or of any units of it must recognize the dual purposes of schools. Although the underlying philosophy and societal demands have not been made explicit, current developments in legislative and other arenas toward requiring greater "accountability" of the schools make the need to recognize the dual purposes clear.

In applying any evaluation principle to education, the purposes of instruction must be clearly recognized. It is true that the purpose of instruction is the modification of behavior. But this must be understood in its broadest sense. Certain segments of the school enterprise are considered to be engaged primarily in modification of specific behaviors. This is chiefly true in so-called skill areas. However, modern communities expect much broader modifications also--those of the type implied in statements such as "ability to get and hold a job," "improved self-image," and "appreciation of cultural differences." These broader objectives clearly imply that instruction must not only result in specific cognitive and psychomotor behavioral changes but it must also operate so as to maintain a physical and emotional environment satisfactory to the community and conducive to achievement of broader objectives.

If the operation of any element of a system (such as an individual employed in it) is to be evaluated, all the factors affecting the operation of that element must be determined and assessed. Two important subsystems impinging on every classroom situation are the support services and the administration of the school.

Support services have two broad purposes. One is the support of instruction by providing facilities, materials, and professional assistance in instructional activities. The second is working directly with students to supply the personal assistance and the environmental conditions expected and required of the schools if their role expectations are to be met.

The term administration is used rather than "school management" since one major function of administration is to mediate between the community, parents, and instructional and support personnel. The other major function is, of course, management per se--the securing and distribution of supplies, equipment, auxiliary services, and physical facilities, and the conduct of coordination and liaison activities.

Since it is a tautology that accountability goes with responsibility, implementation of any system of evaluation of personnel requires, immediately after a determination of the purposes of the organization, a definition of the responsibilities of the various staff positions. In attempting to evaluate certificated school personnel several types of responsibility must be considered. First, there are responsibilities generally devolving upon all such personnel. However, specific statements of the responsibilities of different positions are necessary.

There are many teaching vice-principals, teaching principals, and even teaching superintendents. The custom of using counselors who teach one or more classes is widespread. Department heads or other lower echelon administrative personnel are rarely free from direct instructional duties. It must therefore be recognized that no position title can be assumed to convey the full description of the functions and responsibilities of an individual occupying the position. Valid evaluation of the performance of any individual must be based on an accurate description of the responsibilities assigned to that individual and not on job titles, or even on broad-category statements of job responsibilities--e.g., a general statement of the responsibilities of the position of teacher does not provide sufficient distinction between or definition of the responsibilities of teachers of different student groups, grade levels, and subject areas.

While broad guidelines as to what persons in teaching positions are expected to accomplish are desirable, they cannot constitute an adequate basis for evaluation. They must be supplemented by specific statements as to the results expected. This applies not only to pupil progress but also to managerial functions such as provision of proper environmental conditions and exercise of control over pupil activities. When the authority or resources needed for any function assigned to a teaching position are not under the control of the person in that position, the responsibility must be clearly assigned to another position.

Support responsibilities may be assigned to any position in a school district. Some positions may be exclusively responsible for supporting instruction by working on the psychological and social problems of pupils. Others may be responsible only for furnishing necessary instructional materials. Most will have several responsibilities--e.g., a librarian is typically responsible for providing library materials, for helping students learn to use the library more effectively, and for maintaining order among students who are using the library on an individual basis.

The responsibility of an administrative position can be very broad indeed. However, no person who fills such a position is Superman. Broad statements of responsibility, while theoretically true, are usually unhelpful unless supplemented by specific statements of operational responsibilities. For example, a principal may be said to be responsible for the educational program of his school. Specifically, though, he is responsible for teacher assignments, for maintaining liaison between his staff and the central administration, for seeing that clerical tasks are performed on time, and for supporting the teachers vis-à-vis parents and problem students.

Whatever responsibilities may be assigned to a position, an evaluation of the success with which they have been discharged can only be made in relation to the goals and objectives of the enterprise of which it is a part. In implementing the Stull Act or any other attempt to evaluate the educational enterprise, the nature of the goals and objectives of American education must be considered.

The original goal of American schools, as stated, was to teach children to read the Bible. In the next two centuries, the preparation of persons equipped to be clerks, accountants, and shipmasters was added. Later, the imparting of common social and political goals and the stimulating of a common patriotism were added. These and other goals were not added for schools in all parts of the country, nor were their implications ever analyzed at the local level. They were synthesized into such generalities as the "Seven Cardinal Principles" at the national, academic level.

The natural consequence of this development was the creation of widely varying individual and local interpretations of what the goals of the schools and of any given school program really were. As a result, any school person was forced to enter into a guessing game and to gamble that his idea of the goals he was to strive for would be the ones valued by whoever evaluated his work. To avoid the chaos caused by this situation, which still exists in great degree, educators preferred to evaluate each other--and to be evaluated themselves--according to how they acted, how they appeared, and their experience and academic record.

Research has shown little connection between these presage and process evaluations and the achievement by pupils of any agreed-upon levels of success in any area. This means that evaluation has been a matter of "pleasing the boss" and "keeping your nose clean." The movement toward establishment of objectives which are clearly set forth before school activities commence is an attempt to assure that personnel will be evaluated according to the results they get in progressing toward objectives that are truly relevant to the purposes of the schools.

The terms "goal" and "objective" are frequently used as though they were absolutely distinct. This is not true; there is no precise boundary between goals and objectives. In a pure form a goal is a statement of a purpose of the schools. It gives direction but no specifics. Thus, "to produce better citizens" is a pure goal statement. On the other side, a pure objective is a statement of a specific observable result to be accomplished by a particular educational operation within a specific time span. For example, "by the end of the second week of school to have every pupil write his name in the correct position on each paper he does in class" is a pure objective statement.

Some theorists hold that a sufficient number of pure objective statements in relation to any educational goal will suffice to determine the extent to which the goal is being served by the schools. This is both theoretically and practically fallacious--theoretically, because it does not provide for the interaction among human experiences in developing skills and knowledge which may (and usually does) make the whole quite different from the sum of its parts; practically, because any human behavior, even such a clearly defined one as reading, involves far more specifics than can be noted and recorded within the limits of a school day, a school year, or even the entire school career of a student. Problems of record keeping and staff time only serve to reinforce the impracticality of the pure objective approach.

If pure goal statements are too general, and pure objective statements are too limited, what is the answer? It lies, of course, in some form of intermediate statement, or in the limitation of objectives to be considered in evaluating a school activity to a feasible number, with the hope--unverified by research and contradicted by most experience--that students will somehow fill in the gaps on their own. There is no need to accept the latter suggestion, since there can be a complete range of statements bridging the gaps between those that are "pure objective" and those that are "pure goal."

Because goal statements have typically been too vague for any progress toward the goal to be monitored while instruction is occurring, and because pure objective statements provide for definitive determination of whether or not a specific instructional program has attained the stated objectives, many groups and individuals have chosen to use the latter in evaluation. Unfortunately, many have overlooked or forgotten the limitations of specific objective statements and have denied that there can be intermediates bwtween these and general goal statements. This is not true. In the present state of education only a few specific objective statements can be genuinely useful in evaluation. Intermediate statements are a necessity.

The term "behavioral objective" is particularly unfortunate in connection with evaluation systems. The well-publicized and widely disseminated positions of those espousing specific goal statements have led to a serious misunderstanding among many educators about the nature of objectives. While it is probably true that educational experiences modify the internal psychologies of individuals, even if there is no external manifestation of change, the only way data can be collected for assessing the effects of educational programs is by observation of the behavior of those participating in them. This behavior may be manifest through any type of human activity. Saying certain things in response to a verbal stimulus known as a question is a form of behavior. So is making marks in specific places on a form in response to a visual stimulus. So is spontaneous reading of a book, or refraining from attacking a smaller person.

It is a fact, then, that all statements of objectives which will contribute to the collection of data relevant to how an educational program is succeeding must be behavioral. The behavior may be such that changes (if any) will be very subtle, or it may be so simple and overt that change, or lack thereof, is easily determined. The fact that certain psychologists, and promoters of instructional systems, have attempted to redefine behavior to include only the latter type is to be regretted and does not alter this fact.

In proposing evaluation systems to implement the Stull Act, and in assessing the quality of proposed systems, it is necessary to remember that evaluation is a judgment. If it is based on data relevant to the purpose of the activity being evaluated, it is good evaluation; if it is not based on data, or is based on irrelevant data, it is bad evaluation. Evaluation is also a judgment based on criteria for satisfaction. If the criteria are realistic, it can be good evaluation; if they are not realistic in terms of the situation being evaluated, it will be bad evaluation. If either the purpose of the activity or the criteria are changed in the course of the activity, it will be ex post facto evaluation. Insofar as programs are concerned, being ex post facto does not automatically make evaluation bad. However, if the evaluation is to be used to determine the adequacy of the performance of an individual working in the program or responsible for it, then ex post facto evaluation cannot be validly applied, and its use is an obvious denial of due process to the individuals involved.

The requirements for a satisfactory evaluation system involve several factors. First is the objectives. The purpose of an activity must be determined before any judgment can be made as to whether or not, or to what degree, the activity is succeeding. The essential requirement of an objective is that it be possible to determine by some type of observation whether or not it has been achieved. It is an improvement if it can be made possible to observe partial achievement or "near misses." However, quantification of the degree of achievement of an objective or of the wideness of the "misses" is not essential. A false appearance of objectivity can be given by making a statement such as: "By the end of the semester 80% of the students . . ." The decisions that 80% is the objective and that the particular dimension is relevant are purely judgmental. The use of specific numbers does nothing to make such an objective more useful. This type of quantification is usually more harmful than helpful.

Second is provision for collection of data. Very simple objectives call for collection of only one kind of data. For example, if the objective is "to teach pupils how to get the basketball through the hoop," the only data to be collected is the number of times the ball goes through the hoop. This can be refined by recording also the number of misses, and mathematical manipulation can take place, such as calculating the percentage of tries resulting in success, variation among students in percentage of success, and so on.

Most objectives, however, will require collection of more than one type of data. For example, if the objective were to "teach pupils how to score in a basketball game," we would have to add several dimensions to the one of getting the ball through the hoop. Perhaps one dimension would concern the number of fouls committed, another might concern judgment in passing, another might concern ability to judge when to try for a basket, and so on.

A dimension is an activity or a condition whose quality determines whether or not an objective has been met. Usually this dimension also determines the degree to which the objective has been partially met or the closeness of "misses" for objectives that are not absolute. (An absolute objective is one that is either met or not met and concerning which there is no interest in how close the activity came to meeting it, if it were not met.)

In addition to being logically related to the objective, a dimension must meet the test of practicality. It may be impossible to collect data in a given dimension. For example, an objective might be to "increase student appreciation for American literature." Several dimensions would relate to this objective but the one that might first come to mind--how the students' appreciation actually changed during the course--is one on which no data can be collected (at least not until mind-reading techniques are perfected). Instead, dimensions from which inferences about appreciation of American literature might be drawn must be substituted--e.g., books read, extent of participation in discussion on American literature, opinions expressed, and so on.

Practicality also includes the effort that must be devoted to data gathering and recording. There is a limit to both teacher and student time and, in most cases, to the clerical and other assistance available. Obviously a dimension that greatly escalates the time and resources required for data collection and recording should be considered only in special cases.

Third is criteria. The collection of data for dimensions relevant to an objective is not evaluation. The summation of the data can be termed an assessment of the results of an educational program relative to the objective. Whether the results are to be judged as satisfactory, as unsatisfactory, or as partially satisfactory depends on the evaluative criteria chosen.

Criteria must be unambiguous and unequivocal. They need not be numerical. In fact, the use of numerical criteria may be harmful because they give a false appearance of objectivity to a political and subjective process. Some criteria are clearly indicated by the objectives to which the dimensions relate. For example, if it is accepted that one objective of driver training is to enable students legally to drive motor vehicles, then the criterion for success is clear. The objective has been met for any student who successfully secures a license and has not been met for one who does not get a license.

However, most criteria involve some type of quantitative statement, such as "80% of the students will . . .," "the student will achieve a score of 10 . . .," "the student will have a markedly increased ability to . . ." (The last is a quantitative criterion even though numbers are not used.) The problem in all these cases is to arrive at some type of consensus on what degree of success is acceptable. In the numerical cases, these numbers are simply guesswork, judgments based on experience, or compromises of several such judgments and guesses. In the last example, the definition of "marked increase" or the judgment that the change is a "marked increase" is likewise found to be either a guess, a judgment based on experience (educated guess), or a combination of these and other factors.

Any process whereby guesses, opinions, and judgments are compromised and pooled is a political process. However, in many educational situations the process is more overtly political, with staff politics, governing board politics, and community politics all entering into the establishment of criteria.

Fourth, the validity of an evaluation is not the sum of the validity of its components, but rather the product of these components and usually some extraneous factors. For an evaluation process to be valid, it must be based on valid objectives. It must be based on data collected in dimensions relevant to these objectives. The evaluation criteria must be realistic, related to data, and determined before the educational activity begins.

Extraneous factors apply more to the conduct of educational activities by individuals and small groups. The assessment aspects of evaluation may be applied in any circumstances, but criteria for satisfaction and dissatisfaction relating to the specific performance of the programs and of individuals conducting them cannot be derived for dimensions over which the responsible organizations or individuals have no control. In other words, no success criteria can be validly used in evaluating any educator or educational function on the basis of data in a dimension over which he or it does not have control.

The next step in the evaluation of educational programs is the evaluation of the effectiveness of the services rendered in specific positions. The requirements for valid evaluation have been discussed above. If a particular position carries direct responsibility for the program and carries no other responsibility, evaluation of the program is evaluation of the functioning of that position. Such situations are rare, if they exist at all. They may be approximated in large educational enterprises where marked specialization is possible.

Typically, educational positions have responsibilities for at least some aspects of several programs. Teachers, even those with the most academic assignments, are charged with advising students from time to time, with ordering, or advising on the ordering of, supplies and equipment, and with helping determine general school policy on the application of district policies to such matters as discipline. Principals are usually charged with curriculum coordination and personnel administration in their schools. They typically have responsibilities in other fields such as community relations, coordination with other schools, discipline, and advising the central administration. Central administrators frequently wear so many hats that central office responsibility for various programs is diffuse.

Job definitions are vitally important in the evaluation of the services rendered in any position. As stated earlier, valid evaluation must be based on valid objectives. Insofar as the entire school system is concerned, it may be held that the objectives adopted by the community through the governing board are, if properly stated, automatically valid. Be this as it may, objectives for various positions are valid only if they clearly relate to the responsibilities assigned to that position. As a reductio ad absurdum, it is clear that an objective pertaining to arithmetic computational skills would be invalid for the evaluation of a teacher of English literature. Unfortunately, the relevance of objectives to the evaluation of specific positions is not always so clear.

The general rule for determining the validity of objectives for evaluation of the functioning of a specific position is simple. It is, "Is there a clear, recognized, and formal assignment of responsibility for each aspect of the program to this position?" If the aspect is complex, is there a clear determination of which subdivisions of it belong to this position and which do not? The functioning of the position can only be evaluated when this question has been unequivocally answered for all program elements. Naturally, in most cases the answer for some program elements will be "no." Objectives for these elements will, then, have no relevancy for evaluation of the functioning of the position.

It must be remembered that evaluation of the functioning of a position is necessary for the evaluation of personnel assigned to the position. But, to use a mathematical metaphor, it is not "sufficient." First, the official assignment of personnel must clearly state that this position is one to which the individual is assigned. In the case of full-time positions this may be simple, but there are many positions manned by one or more part-time persons. This is frequently true in counseling and consultant positions.

Second, and of vital importance, no individual can be validly evaluated in a dimension over which he has no control. Teachers cannot be evaluated on the attainment of objectives where text and other materials are required for instruction if they have no control over the provision of these materials. Teachers cannot be evaluated on an objective relating to student social behavior if they have no control of disciplinary processes, interrelations with parents, and other elements relevant to such an objective. Similarly, principals cannot be evaluated on the basis of their success in getting supplies to teachers if they have no control over the allotment and distribution of supplies.

To summarize we can say that, to be fair and to be based on due process, evaluation requires establishment of goals and of objectives specific enough to define dimensions on which data can be gathered. Very narrow goals will generally not make possible an evaluation of the program as a whole. Numerical data are not necessary for valid assessment of educational programs. Criteria for success are derived from guesses, informed judgments, and compromise. In no case can they be considered to be absolutes.

In short, evaluation is the application of agreed-upon criteria to the data collected for dimensions relevant to the objectives adopted for the educational unit being evaluated.

To apply the results of an evaluation to a given position, it is essential that the responsibilities of that position in relation to the objectives be unequivocally defined. Where individuals holding positions are to be evaluated, the positions description must be accompanied by an equivocal assignment of authority plus an unequivocal statement as to which other personnel are responsible for those areas where no authority is given.

Individual evaluation can validly be accomplished only by referring the evaluation of programs to the positions responsible, and then applying the results only to those factors for which the individual has both responsibility and necessary authority to act. Authority does not only mean authority over personnel; it also includes authority to secure supplies, modify programs, discipline students, and initiate joint activities with other personnel.

Guidelines

On the basis of the considerations discussed above, the California Teachers Association has arrived at a set of guidelines for considering the comprehensiveness and quality of district proposals for implementing the Stull Act. Since there are many ways to take the various factors into account, and since the desirability or undesirability of using a particular process can only be determined by experience, these guidelines are primarily directed toward gaining consideration of all necessary factors rather than toward promoting specific ways of coping with the factors. The guidelines now in use are as follows:

1. Standards for Pupil Progress

1.1 Have goals related to each certificated employee assignment been established and clearly stated?

General Considerations. It is essential that there be a formal statement of the specific goals for a given class or activity. Although they may be subject to change from time to time, these goals should, in the aggregate, constitute a form of mission statement for the activity.

The nature and degree of responsibility of each certificated employee for the various segments of the mission must be specified.

Goals and responsibilities must be unambiguously stated with particular attention to the avoidance of overlapping.

Scope. The goals statements should not try to cover the entire program, particularly in early stages. On the other hand, there should be provision for assuring that no important goal be left out.

Specificity. While goals must be stated more specifically than most have been in the past, care should be taken to avoid defining them so narrowly as to make them mere objectives.

Organization. In addition to assuring adequate coverage and avoidance of overlapping, goals should be examined to determine their consistency. If relatively independent goals are stated, provision must be made for determining priorities.

1.2 Have pupil-progress objectives for each organizational unit of the school system, such as level, site, class, function (e.g., pupil personnel services), been included in formally adopted statements?

General Considerations. Objectives are more refined and narrowed statements of goals. They must be consistent with and relevant to the adopted goals. They must include some statement of the type of progress or change expected, and this progress or change must be of an observable nature.

Scope. Most educational activities are intended to produce results which are extremely complex. Any attempt to specify all possible objectives is self-defeating because of the immense amount of observation, other data gathering, and record-keeping involved. Resources of pupil as well as staff time are not adequate for a comprehensive development of, and assessment of progress toward, completely specific objectives.

Specificity. Extreme specificity of objectives is to be avoided. Only a relatively few can be stated and used in evaluation. The more specific the objectives used, the wider the gaps left. On the other hand, an objective should be more specific than a goal and should refer to concrete results capable of evaluation.

1.3 Have success criteria statements for each objective been formally adopted?

General Considerations. Objectives, being statements of concrete results desired at the conclusion of a given activity or phase of an activity, must contain statements of how much such concrete results will be recognized. This recognition may result from gathering data by use of testing devices. It may also result from observation of student activities, study of records relating to school and district accomplishments, or professional perceptions.

Nature of Criteria. Specific numerical criteria for success can validly be set for few educational objectives. Such criteria should be carefully examined for relevance and validity. "On-off" or "go-no-go" criteria are appropriate for many specialized activities--e.g., a driver training student either can or cannot pilot a car successfully. Most criteria must involve multistage definitions of success. Partial success must be recognized.

2. Process for Evaluations Relating to Pupil Progress

2.1 Is there a formally adopted statement of methods to be used in assessing pupil status in relation to each success criterion?

Assessment of pupil achievement or changes in this achievement (pupil progress) requires the collection of data which may come from many sources. Ability of pupils to give or mark answers generally accepted as correct on paper-and-pencil tests is one source of data. Ability to perform specified tasks in other types of activity may be another source. Frequency of occurrence of specified types of behavior either of individuals or groups (e.g., classes) is another. General impressions leading to considered judgments of professional personnel are still another source of data.

2.2 Has an adequate mechanism been established for assuring valid application of each method of assessment of pupil status?

General Considerations. Paper-and-pencil tests must be relevant to the goals of the instructional activity. If commercial tests are used, items must be checked for such relevance and those that do not meet this requirement must be eliminated from consideration in assessing pupil achievement. This restriction applies to items in performance and other tests that are not the paper-and-pencil type.

Provision for Collecting Data. Obviously, needed test materials must be made available. This includes provision of time and technical assistance in preparing locally made supplementary tests. Where observation and professional judgment are used as sources of data, adequate training in observational techniques and record keeping and adequate time and clerical assistance for these activities are required.

Uniformity. When various segments of an educational activity, for example different classes in the same subject at the same grade level, are to be evaluated, a uniform procedure for collecting and interpreting data is required. There must be sufficiently definite, mutually acceptable guidelines to assure this. Their specificity may be increased as experience in their use accumulates. Of course, adequate time and consultative and clerical assistance must be provided.

3. Are There Formally Adopted Procedures and Criteria for Relating Environmental (Teaching) Conditions to Evaluation of Success in Terms of Pupil Progress?

   3.1 Is there a formally adopted statement or list of environmental conditions to be considered in evaluation of staff on the basis of pupil progress?

General Considerations. This is one of the most crucial aspects of an evaluation policy. Criteria should be divided into three categories. This division will to some degree dictate a division of procedures into three parallel categories.

Classroom Criteria. Classroom criteria should be quite specific. The basic principle is that control must be adequate to (a) ensure the possibility of accepted learning activities, (b) prevent any pupil or group of pupils from harming other pupils or preventing them from carrying out reasonable activities, (c) ensure fair and orderly distribution of necessary supplies and learning aids, (d) arrange and carry out appropriate activities for securing special assistance and services where needed by pupils (either individually or in groups), and (e) prevent interference with proper activities of other classroom units or with the general operation of the school.

Nonprofessional Administrative Criteria. These, too, should be specific. The basic principle is that service must be adequate to ensure (a) adequate physical facilities (both structure and equipment), (b) timely and adequate supply of needed materials and equipment without excessive red tape, (c) reasonable pupil loads, either in terms of class size or number of pupils for which a service is responsible, and (d) reasonable amounts of secretarial, clerical, and other nonprofessional services to enable professional personnel to perform their professional tasks.

Professional Administrative Criteria. In some areas these criteria should be specific. The criteria should ensure (a) proper assignment of personnel both as to nature of assignment and as to total responsibility load, (b) administrative support for all personnel in enforcing school and district regulations, (c) administrative support and assistance in securing public and parental cooperation in the treatment of pupils with problems, (d) adequate and relevant in-service training, and (e) professional participation of certificated staff in the development and modification of educational programs. In other areas, professional administrative criteria may need to be more general. These general criteria include, but are not limited to (a) support, with the governing board and public of reasonable requests of the staff, (b) promoting citizen understanding of school needs and policies, (c) opposing unwarranted attacks on the staff, both individual and collective, and (d) accepting or supporting no innovative or other proposal without providing for voluntary staff participation and objective evaluation or results.

3.2 Is there a formally adopted procedure and approved method of assessing environmental conditions?

General Considerations. Data for assessing environmental conditions and level of control will derive from several sources--reports (both good and bad) of pupils and parents, observations by staff members, observations by the individual responsible for a given function, district and school records, and observations of outside agencies or persons.

Classroom. Since complaining pupils and parents (and sometimes commending ones) are common, a specific procedure is needed for handling such complaints and commendations when made officially to a person other than the individual complained about or commended. Provision must be made for evaluating such complaints and commendations. Only those described in detail in writing should enter any personnel file for purposes of evaluation. Complaints or commendations from fellow staff should be handled in the same way, regardless of the position of either party involved.

Administrative. Records of requests for service, appropriateness of such requests, and utilization of services furnished should be kept if any evaluation is to be made in this area. Observation by teachers and others, including outside evaluators, should be used in determining physical environment. Both observation and district and school records will furnish data on adequacy of facilities, equipment, and supplies. Records of class size, of disciplinary and other actions supporting teaching and other personnel, and records of the nature and size of helping staff will all furnish data for evaluating the professional and nonprofessional aspects of administrative services.

Professional Judgment. People who request administrative support, supplies, and other needed items to facilitate instruction are in a very good position to assess the adequacy of most environmental factors. The regular collection of data on their judgments on these factors is important in securing comprehensive information on administrative and other supporting services.

3.3 Is there an explicit statement of criteria for determining adequacy of environmental conditions in relation to pupil progress criteria?

3.3.1. Does the statement include minimum standards to be met in environmental conditions? (These are standards which, if not attained, are recognized as relieving the person working under substandard conditions from all responsibility for satisfactory pupil progress.) Minimum standards for acceptance of responsibility should be set realistically. Some that should be specified are: maximum class size (which may vary by grade level, subject area, and types of pupils); physical facilities adequate to provide stations for all assigned pupils; conditions which would be hazardous to health (e.g., ventilation failures); maximum noise levels from external or unavoidable sources; minimum supplies and equipment for required activities; and delivery of supplies (texts, aids, materials) in sufficient time to permit their use in required activities.

3.3.2. Does the statement include formally adopted procedures for relating environmental measures to pupil progress measures in assessing staff performance when the minimum standards have been met, but conditions are not perfect? The standards for borderline conditions are more variable but cover at least all the items listed for minimum conditions. The weighting to be given to the various items and the spread between criteria for minimum and fully satisfactory conditions must be arrived at by consensus in each situation. Provision must be made for developing and from time to time revising these standards.

4. Are There Formally Adopted Guidelines for Extra-Duty Assignments?

General Considerations. Extra-duty assignments, or "adjunct duties," should be few and should involve activities that cut across several organizational units. In some cases, where no organizational provision is made for certain activities--e.g., student government, interest groups, playground control--duties relating to these may be assigned to persons whose regular responsibilities do not relate directly to these activities. Where possible these duties should be formally assigned and accepted as extra assignments with the clear understanding that performance in this type of assignment will be evaluated separately from general evaluation of performance.

4.1 Are there procedures for determining the load represented by such assignments?

Since adjunct duties may not always be avoided, and since the work involved reduces time and energy available for activities required by the primary assignment, some procedure for determining the load represented by adjunct duties must be developed. This may take the form of developing an agreed-upon list of such assignments, others ranked according to load, or grouped by load categories. More elaborate procedures involving weighting factors and formulas based on clock hours required, pupil contracts, and total duration of an assignment may be developed. Both methods have been used in school districts. Studies on policies relating to "extra pay for extra duties" give examples of the use of these methods.

4.2 Is there a formally adopted statement of objectives and of success criteria for extra duty assignments? (Ultimately each such assignment will require a separate statement.)

Objectives and success criteria for adjunct assignments should be kept to a minimum, at least until such criteria are developed for regular assignments. In any case, these criteria will usually be simpler than those for regular assignments. For example, the objective of the coach is to win as many games as possible, with as few injuries as possible, and with no open dissension on the squad. Success in attaining such objectives is easily determined.

4.3 Are there formally adopted procedures for assessing staff performance in extra assignments?

If adjunct duties are defined as extra duties, the assessment of performance will usually depend upon rather simple measurements. Examples are: numbers of students dropping out of an activity, audience judgment of a show put on by a club, member (also possibly parental) satisfaction about participation, and attendance at PTA meetings. These measures need to be formally adopted, however. This means there must be agreement as to what will be assessed (measured, observed, judged) and what criteria will need to be met for satisfactory performance. This agreement should be in writing but, if consistent with general board policy, memoranda or assignment notices can probably contain what is necessary.

5. Proper Control and Suitable Learning Environment

General Considerations. The first two questions below were covered earlier. The major requirements of an acceptable evaluation policy are that the definitions and procedures called for do exist and that, if they are used as an element in evaluation of certificated personnel, specific reference be made to the elements discussed under the section on formal adoption of environmental assessment criteria as the basis for assessing performance.

5.1 Are there formal and specific definitions of "proper control" and "suitable learning environment" for each organizational unit of the school system--district, school, department, grade, class, etc.?

5.2 Are there formally adopted procedures for assessing the degree of proper control and suitable learning environment as defined?

5.3 Are there specific definitions of unacceptable levels of attainment of proper control and suitable learning environment as defined?

Administrative Personnel. Insofar as administrative and supervisory personnel are concerned, the definition of unacceptable levels of attainment may coincide with the minimum standards adopted for environmental conditions. However, it may be that some level of acceptability in the range between minimum and fully satisfactory may be adopted as a criterion. Provision must be made in either case to assure that any item on which personnel are evaluated falls under the effective control of those personnel. To the extent it does not, this item must be treated as an exterior "environmental condition." Procedures for determining these cases and assessing their weight in evaluation of certificated personnel can probably be the same as those developed for assessing environmental conditions.

Teaching Personnel. Standards of classroom behavior should be broadly enough stated to permit allowance for individual teaching styles and differing pupil populations. In all cases the test of adequacy should be whether or not classroom conditions under the control of the teacher, or teaching group, interfere with the attainment of the goals of the educational activity involved. This same test applies for determining environmental adequacy. As for administrative personnel, procedures must be provided for determining which factors of control and environment are under the effective jurisdiction of the teaching staff involved, and for relating the state of these factors to any elements used in evaluating such personnel in connection with control and environment.

6. Are There Procedures for Improving, Updating, and Validating Evaluation?

General Considerations. Procedures for review of criteria, assessment procedures, and definitions should, as a minimum, provide for continuous monitoring of the evaluation system with a periodic deadline for submission of a report on its operation and recommendations for needed changes. Data collection should involve all personnel and probably, for some items, students and lay members of the community as well. There should be a provision for a continuing group to classify, coordinate, and evaluate these data and to develop proposals for change. This should be a professional group with representation from all personnel classes. A significant portion of nonadministrative personnel should formally represent the local association of the district. Inclusion of a representative of the governing board is optional. Out-of-district personnel should be used only as consultants.

6.1 Is there a formally adopted procedure for review of objectives, success criteria, and assessment procedures?

6.2 Are there formally adopted procedures for review of assessment and criteria relating to environmental conditions?

6.3 Is there a formally adopted procedure for review and reassessment of fundamental definitions relating to pupil progress, environmental conditions, extra assignments, and proper control and suitable learning environment?

The California Teachers Association has examined approximately 100 proposals for implementing the Stull Act on the basis of these guidelines. None has fulfilled all of the conditions required. Most have failed in all but one or two respects. A considerable number do not satisfy even one criterion. It is obvious that present activities represent a mixture of ignorance, attempts to comply pro forma with no actual change or development in practice, and outright rejection of the Stull Act. There is, however, evidence of serious intent to comply. By this time next year an evaluation of the evaluation systems designated to implement the Stull Act should be fruitful.

DISCUSSION OF PAPER BY GARFORD G. GORDON

JAMES F. GARVEY
Glendora (California) Unified School District

Remember back in 1965 when our legislature enacted Senate Bill 1? We were all elated--we were ecstatic! We even gave it a nickname--the Magna Carta Bill of Education. It was going to change public school education in California significantly because it was going to give local control to each of the school districts. Here was the legislature giving away power to the people, and what did the people do when they were tossed the ball? You know what happened--they fumbled it. We did a little bit, of course--we did take out some of the mandated courses that were prescribed in the Education Code and we did broaden our curriculum by including some electives. And that was the beginning of many new courses in some high schools. But did we effect a significant change in the public schools of California? No. So I raise a question I believe we will ask five years from now: "Whatever happened to the Stull Act?"

I am inclined to think the same thing that happened to SB 1 will occur. I think we are inadequately trained to serve in a leadership capacity in preparing teachers to write worthy and well-written objectives. Also, I don't think the Legislature has provided us with the necessary resources for 1972 that we had in 1971. Initially, the Stull Bill called for $400,000 for in-service training, but that provision was wiped out. Boards are reluctant to give released time to administrators and teachers for the purpose of implementing the Stull Act because released time costs money. So the Legislature did not provide us with the resources, and until it does, I don't think any noticeable change will occur in the public schools of California as a result of the Stull Act.

I predict there won't be an increase or a decrease in the number of dismissals on March 15, 1973. I also predict that there won't be a significant improvement in the quality of instruction during the 1972-1973 school year, and principals won't spend noticeably more time in the classrooms.

You are doubtless thinking that I am an opponent rather than a proponent of the Stull Act. Let me clarify my position. I think there is some merit in the Act; I'm only saying that it is going to be extremely difficult, indeed almost impossible, to implement. For example, how do we answer such questions from a teacher as: "How many objectives do I have to write?" "How do I know whether to put 70% or 80% in my objectives when we talk about level of achievement?" "What happens if I don't write any objectives?" I don't know the answers to such questions, but I continue to be asked them.

We are advising our teachers to make an effort to write the objectives and we will give them all the help, all the assistance, and all the released time we can possibly afford. In fact, we are saying, "You teachers develop the evaluative system for our school district." By law we must develop, and the school board must adopt, an evaluative system. So in my district our Stull Act committee consisted entirely of teachers. Now we have an evaluative system, and it has been adopted. We are in the process of writing our instructional objectives, the principals are writing their school objectives, and the central office administrators are writing their district objectives.

There is an aura of mutuality that pervades this process. However, we are also advising our teachers, when they write objectives, to do two things. First (this sounds like heresy), we advise them to accompany each objective with all the constraints or variables which may impede its attainment. Second, we advise them not to write any objective which they can't fully control. I think that is being realistic and reasonable, and that is probably why we are not meeting with apathy or having confrontations.

There is an effort to go from artistry to some kind of a rigid, scientific method of evaluating human beings; I am reminded, however, of the adage that under carefully controlled conditions the human being still does as he pleases. I am also reminded of Professor Stainbrook's comment about five years ago: "We put children in crowds and we call the crowds classrooms." We did that last year, and we are going to do it again next year. This is one complaint of classroom teachers, although research shows that decreasing class sizes really doesn't result in more effective learning. But the Stull Act, I think, will fall by the wayside, unless in this session (and this is most unlikely) the Legislature comes forth with sufficient money to provide all the resources necessary to implement the Act properly. And that's not going to happen in this session of the Legislature, is it?

One of the pitfalls in implementing the Act seems to be that at the district level we haven't really addressed ourselves to support services. Teachers are hungry for proper support services and we can't provide them. The principal simply shrugs his shoulders and says, "I'm sorry about that--we are a low-level school district," or gives some other answer similar to that.

To sum up, although I think there is some merit in the Stull Act, I don't think it is going to fly, because it wasn't accompanied with sufficient money to provide efficacious in-service training programs for all of our certificated people and the number of dollars required to provide the support services.

ROBERT L. FORBES
Castro Valley (California) Unified School District

I had the pleasure of reading Dr. Gordon's paper and I am delighted as a CTA member as well as a member of the Association of School Administrators that he is working for me, as I think his research is interestingly done and I know that when the time comes and we achieve perfection in the educational world here in California, he will have a hand in getting us there. But I do feel, along with some of you, that maybe the paper told me more about tigers and penguins than I need to know at the moment.

I have some problems with the Stull Act too, because I am on the firing line in trying to get it to work. And I couldn't agree more that it should have been funded in terms of released time, in-service education, and so forth. However, I think we also must realize the motivation that lies behind the Act. I think the question in California really concerns who should be educated and for how long. When we look at the level of state support in California, for public schools relative to our per capita wealth, and when we hear all the groaning from the legislature and the Governor's office about how much education is costing, we realize that this is a political expression of frustration about what is wrong with the society and how to blame the schools for not doing the job they are supposed to do.

This feeling of frustration started back with Sputnik, and we still have it. My friends in the business community have a very difficult time relating to the question when I say to them: "Well, we did very well what you wanted us to do to prepare people for the world of work--now where are you with your jobs?" And from that point on the dialogue goes down hill very fast.

Another thing bothers me somewhat. Why, in a state as sophisticated as California, didn't they select a few pilot schools to work with the claims of criteria that are being related to you here at this Conference before they jumped in with both feet and got the whole state embroiled in this situation? If the motivation was really to improve education and to improve teacher and administrative performance, this seems a logical first step to have taken.

But the law is the law, and all the quibbling about it does not get away from the fact that school districts must try to comply with the law. The question is, how do we comply with it in such a way as to maintain our integrity, maintain good teaching professional and administrative relationships going, and perhaps learn something in the process that may be of use? It is not really a question of blind obedience to a poor law, or moaning and groaning about the cost of it, and so forth. It is a question really of taking a look at what it might be possible to do and establishing a working hypothesis that something can be done if we continually evaluate it and try to make it a little better as we go along. This is precisely what we did in the Castro Valley Unified School District. We have a procedure for Stull Act implementation that I'm sure Dr. Gordon would find is full of holes. It probably has as many holes as the proverbial Swiss cheese, but it also has some people nibbling and maybe getting some nourishment.

I think we must get a few things in perspective. It seems to me that the school districts that didn't do very much with SB 1 (and I'm glad the previous discussant mentioned that) are going to have some problems. Our plan for implementing the Stull Act is predicated on the fact that in the past couple of years we have involved our community in the need assessment of its students and then followed this up with some work on behavioral objectives. But districts who didn't do any of this and now are faced with the Stull Act, which is in many ways predicated on some work on those areas, have a lot of catching up to do.

Let's assume then, for the sake of this short discussion, that we did some work with Senate Bill 1 and we prepared some goals and objectives. We did some student need inventories, as imperfect as they were, and we got down to about level 5 on behavioral objectives last year. At the same time, we formed a committee concerned with the Stull Act. We went to the president of the District Teachers Association and asked for some people from his group to sit in on a working committee. We secured some material from the CTA, and the M.A.P. papers from ACSA and we sat down one day after school and said: "What can we do?" We had three or four administrators, eight or nine teachers representing every level from kindergarten through 12th grade, representatives of different administrative positions, and so forth. We looked at each other and said: "What do we do?" Well, the first thing we did was a lot of listening, because there isn't a person who doesn't have some kind of position about the Stull Act and the philosophic and educational implications behind it.

This would be my first advice to you--listen to everybody tell of his educational frustrations and, if you trust what you are listening to in terms of the people who are doing the talking, you have taken the first step on putting your district on the road to the possible use of the Stull Act. I can't emphasize this too strongly--you must trust, accept, and be professionally and personally involved with the people and their positions, whether you agree with their positions or not. If you don't have this kind of rapport in your school district, it's going to be more difficult to get this Act implemented.

Our committee had about three talk sessions before we began to act. We wrote up a series of procedures for people to take back to their schools to get feedback from their colleagues. In a couple of months we came up with a rough draft of a first decision paper. The superintendent called the entire certificated staff together in one of the schools (and again it wasn't on released time, because we didn't have it). Each member of the committee, representing different points of view--administrative, teaching subject matter, grade level--

went through what he or she thought was possible to do for the year which started last July 1. We got more feedback and we went back to the drafting board to write new observation forms, evaluation forms, etc. Finally we were able to report back to our people again through building meetings and the other avenues available in any school district. I should add that we had job descriptions written by the people who were doing the jobs, for all teachers and administrators. The committee made a presentation to the superintendent's administrative council after we had presented our work to the teachers and then we went on to the school board and said: "This is how we intend to comply with the Stull Act for this first year."

One of the key issues in our Stull procedure is the use of a chosen evaluator. We are encouraging people to use any evaluator they wish, with the limitation that nontenured people of course must be evaluated and that the evaluator must be the appropriate principal. But tenured people in the 50% to be evaluated this year can choose any person they wish. The law says "any person"--it doesn't say "certificated person." But what we have going on is, for example, that an English teacher is being evaluated by a vice-principal, a business education teacher is being evaluated by a counselor who happens to have a very good preparation in business education, and so on.

Some may argue that this will lead to "one hand washing the other." But if our opinion of our colleagues is so low that we feel they will simply do the least they can get away with and therefore will look for ways to scratch each others' backs, that atmosphere permeates the entire school district. On the other hand, if we trust people and work with them, if we know who needs help and who doesn't, as well as what kind of help and how to present that help in a positive, nonthreatening way, I think we will realize a lot of growth.

I would also remark that according to the Act, every procedure, every form, and every technique used to meet the requirements of the Act must be approved by the County Counsel. Our procedure did go to the County Counsel and was approved by him with the remark that peer evaluation is very interesting--we may find ourselves in difficulty, but it's certainly legal. I can tell you, after the first week or so of getting people in with their chosen evaluators and going over with them in a three-way conference what their aims and objectives are, that it seems to be working, at least in the sense that people are beginning to know for the first time what their colleagues are doing.

We have been provided a small amount (some $2,000 worth) of substitute time for the evaluators to give them some time free to do the observations of the tenured staff. Another thing we did for the first year was to take an arbitrary stand, as a committee (and the County Counsel approved it), that we couldn't possibly do the kinds of things that Dr. Gordon is saying we should do. I think that some time in the future we must go his way, but at present we could only do what we are trying to do, which is to say: "Look, this is a hypothesis we are going to test, and we will have the same standing committee meet a couple of times a year to see how things are going." We expect to try about two performance criteria for each of two classes or two subjects on the suitable learning environment and expected student progress, adjunct duties, etc. These are two important things we would do anyway, but we are laying them out to see what kind of help we need to do them and how we expect to accomplish them.

We haven't had very much mumbling and grumbling. I won't say that each of our staff members in a district of 7,000 students is ecstatic about the Stull Act. Of course they are not, and neither are the administrators. But we have a positive thing going; people understand what we are trying to do, and they are giving us a great deal of cooperation. I think we will probably see some benefit.

In conclusion, I must dispute a little the statement of the superintendent who preceded me. As a building principal I have already been in classrooms more than I was a year ago or two years ago at this time, and basically because of the Stull Act. I have had to check out some of the things people said they wanted to do, and I felt the best way to do it was to observe classroom situations before personnel came to me with their objectives for expected student progress. I agree it is a headache. I think the motivation the legislature had is wrong, the funding arrangement is wrong, and the lack of pilot schools is wrong. But I think we can comply with the Act, given the caveat that I must give as a former social science professor that the multivariate nature of human behavior probably defies measurement anyway. I couldn't agree with Dr. Gordon more on that. The variables involved are fantastically difficult to measure with the tools we have. Even if we could measure them, I am not so sure, for example, that I want to know what all the component parts are of a rose or sunset, and to destroy the beauty of that artistic creation. But I think there is a middle ground somewhere that we can reach by acting on the principle that maybe there is something we can learn by trying this process as long as we work knowing what the limitations are, and more than that, as long as we work at the time time in every way we can make certain the legislature understands that we don't appreciate its distrust. We haven't earned it, and we don't deserve it.

OBEYING AND IMPROVING THE MANDATE TO EVALUATE EDUCATORS: SOME LESSONS FROM THE CONFERENCE ON THE STULL ACT

N. L. GAGE
Stanford University

What lessons have emerged from the conference's intensive consideration of the Stull Act? The proceedings include contributions from experts on law, measurement, statistics, educational objectives, teacher education, administration, accountability, and the organized teaching profession. In what follows, I try to bring together some lessons from the papers, discussions, and comments from the floor. My remarks reflect only one person's impressions and interpretations of some salient ideas. They should not be regarded as a summary, because the various contributions are too rich to permit any substitute for a careful reading. Rather, this personal distillation should serve only as an aid to the reader in marshalling his own ideas and conclusions.

My statements take the form of Don't's followed by Do's, with some elaboration of the rationale underlying them. This style is intended to sharpen the issues. The sequence of ideas is one of increasing controversiality. They start from what looks like sure ground and move from there onto thin ice. At the end, I shall look at ways in which the State Legislature and the Governor could substantially improve our chances of achieving the goal of the Stull Act--better education for California's children and youth.

1. Don't specify objectives in minute detail and great numbers. Rather, specify objectives only to the degree that

perhaps 20 objectives are identified in each subject for each semester.

There seemed to be consensus on this point among all those who touched upon the subject of educational objectives in the conference. Advocates and opponents of behavioral objectives agree that spelling out hundreds of behaviors to be achieved by the students in a course tends to subvert the purpose of the enterprise. As Popham put it, "Care should be taken when adopting objectives as a descriptive vehicle not to generate so many objectives statements that the result is an intellectually unmanageable collection.... Although no empirical studies have been conducted on this point, one suspects that being able to cognitively process many more than 15 or 20 objectives per academic year is quite unlikely, and that that number probably represents sufficient detail for most educators" (pp. 123-124).

As was true of many other issues considered in this conference, the touchstone is reasonableness. Any approach to compliance with the Stull Act can be reduced to absurdity if it is pushed to an extreme. If it is true that not all educational goals and objectives can or should be specified in behavioral terms, it does not follow that none should be stated in that way. If it is true that extreme specificity and large numbers of objectives make unmanageable the whole approach of stating objectives as a guide to instruction and evaluation, it does not follow that reasonable specificity is unattainable.

2. Don't require each teacher or school to originate all of the objectives at which the educational effort will be aimed. Rather, make judicious use of previous formulations of objectives, such as those available from curriculum

specialists and organizations for sharing objectives, modifying them to meet the needs of specific classes and students.

The rationale here is also clear: Since many educators have already produced statements of the kinds of behaviors that they want their students to acquire, teachers should not have to reformulate such statements from the beginning. If teachers and administrators make reasonable use of the statements of objectives already available, the task of formulating such objectives can be substantially cut down. Such use means, of course, that objectives will be modified, as judged appropriate, to suit local values and needs and abilities of one's own students. One hears stories about thousands of teachers devoting hundreds of hours to the task of writing educational objectives. These stories suggest that the approach recommended here is not currently being taken. If it were, much effort might be saved without loss, and even with some gain, in the quality of effort needed to comply with the Stull Act.

3. Don't expect objectives in all school subjects to be set at the same level of uniformity from one teacher to another, regardless of differences in the amount of expert consensus in the different subjects. Rather, expect more uniform objectives in some subjects (e.g., reading and arithmetic) than in others (e.g., art and history).

The degree of consensus about what students ought to learn varies from one subject matter area to another. Although this dimension of curriculum has not been measured objectively and quantitatively, it exists and ought to be taken into account. Outsiders probably perceive more agreement on objectives than do the insiders who work on the curriculum in a given subject matter area. But even the specialists within the various areas will probably agree that there is greater consensus about what students ought to learn in some subject matters than in others. If so, teachers ought not to be expected to agree as much with one another on the objectives they set forth for subjects like history, English, or even science, as they would for subjects like reading and arithmetic. The degree of consensus may also be greater in the lower grade levels than the higher ones. Again, educators ought to avoid insisting on very high or very low agreement among the objectives set forth by various teachers.

4. Don't concentrate exclusively on student achievement of cognitive, or intellectual, objectives. Take into consideration student achievement of attitudinal objectives, including attitudes toward the teacher and the subject matter taught by that teacher.

This recommendation seems to be unarguable, but it is worth stating because of the degree to which discussion has focused almost exclusively on cognitive objectives and has neglected the goals of helping students to develop desirable attitudes toward learning. The importance of such attitudes is not a question to be settled by empirical research. Rather, such attitudes are important because of society's values. We should take into account how our students feel about their teachers and the subjects taught by their teachers, regardless of how well they perform on tests of their achievement in those subjects. Teachers whose students score well on tests but nonetheless have little positive regard for their teachers ought to be helped to remedy an undesirable situation.

5. Don't insist that all pupils from all social, cultural, and ethnic backgrounds meet the same standards of

achievement of cognitive objectives. Rather, recognize the educational validity of different standards of progress for pupils from different backgrounds.

Professors Alkin and Baker noted that "the potential political consequence of publicly assenting to different standards for different student groups could be devastating, even when variations in antecedent experiences are clearly documented" (page 249). Nonetheless, such public assent is necessary. It is a fact that students from low-income areas do less well on achievement tests than those from middle-income areas. No survey of student achievement of the last fifty years, including such massive enterprises as Project TALENT and the Equal Educational Opportunity Survey, which resulted in the Coleman Report, has failed to reveal such differences. This is not to say that we should rest content with these conditions of educational achievement. But they will not be abolished by any fiat. This means that setting expected standards of pupil progress that are the same for all students, regardless of the income areas from which they come, will inevitably subject some teachers, schools, and school districts to unfavorable evaluations.

Instead, standards must be adjusted to reflect what can reasonably be expected in terms of student achievement in different socioeconomic areas. The mechanisms for setting such expectations need to be subjected to research and development so as to reflect not merely what can be predicted statistically on the basis of students' prior achievement and characteristics but also, as determined in more detailed and perhaps clinical fashion, what is appropriate for each teacher individually.

6. Don't use students' scores on standardized achievement tests to evaluate teachers. Rather, use students' scores on teacher-made criterion-referenced tests.

The shortcomings of standardized achievement tests for the purposes of evaluating certificated personnel in accordance with the Stull Act have been well set forth by Professors Glass and Popham. Part of the argument against such use of tests is that we should not evaluate teachers in relation to other teachers. We have long known about the problems in evaluating teachers on the basis of their students' achievement. One textbook made this point some 30 years ago:

> [One] set of difficulties in evaluating [teachers] through pupil changes involves the problem of insuring that any difference or changes observed are due solely to the influence of the teacher rather than to other influences.... It is because of the difficulty of taking into account all these...other factors that the approach to teacher evaluation in terms of desirable changes in pupils cannot readily be used...by school administrators....*

These difficulties would be greatly lessened if we evaluated teachers, not in relation to one another, but in relation to the achievement of their own students. That achievement, in turn, should be evaluated in relation to reasonable standards of pupil progress toward reasonable objectives. Comparing any sizable number of teachers with one another means that half of the teachers will by definition be found to be below average. It also means that teachers are pitted against one another in a kind of competition with the result that, no matter how well (or poorly) individual teachers perform, some must inevitably be regarded as going to the top, and others as falling to the bottom. Instead, we should establish a rationale that permits all teachers to be acceptable and doing well; or no teachers; or some fraction in between.

---

*H. H. Remmers and N. L. Gage, Educational measurement and evaluation. (New York: Harper, 1943), pp. 454-455.

The argument being made is similar to that for criterion-referenced evaluation of student achievement. For too long, we regarded the function of achievement tests as that of discriminating among students and placing them in a rank order with respect to achievement. In recent years, we have begun to realize the error of this conception. We now recognize that achievement tests should be constructed so as to reveal achievement of important objectives and that such tests should be evaluated in terms of the significance of their content, regardless of the shape of the frequency distribution of scores on the tests. In fact, the goal should be to teach so that all students achieve maximally, i.e., attain mastery of as many appropriate objectives as possible. When this happens, the distributions of scores on achievement tests will no longer resemble the bell-shaped curve that used to be regarded as desirable. Instead, the distributions will show a bunching of students at the high end of the scale, with only a few students falling lower.

The same approach to teacher evaluation ought to be adopted. If it is, many of the objections to the use of student achievement in evaluating teachers will no longer hold. What we need, if the criterion-referenced conception is applied, is a basis for estimating the degree to which it can safely be said that a teacher has achieved an acceptable level in relation to a criterion of performance, regardless of how well or poorly other teachers have done.

This suggestion does not conflict with that made in my second "Don't." Teachers should make use of existing collections of objectives. But the construction of tests appropriate to what is being taught by a particular teacher in a given subject in a specific school district ought to be strongly influenced by that teacher.

7. Don't assume that student achievement, performance, and behavior--whether cognitive or social-emotional--can serve as the only criteria for evaluating teachers. Rather, use observations and ratings of teachers on validated behavioral dimensions in addition to measures of student progress.

The dimensions of teacher behavior that should be evaluated can be identified from the many studies of the correlations between teaching behaviors and student achievement. Many of these studies have recently been reviewed and summarized by Barak Rosenshine[*]. Although much of that research is vulnerable to sophisticated methodological criticism,[†] my impression is that the trend of the results can be used to select better teacher-behavior dimensions than would otherwise be possible. Those dimensions should be incorporated into observation instruments that would permit trained observers to determine the frequency and quality of the teacher's behaviors. Some teachers may completely lack some kinds of behavior (e.g., teacher approval, teacher acceptance of students' ideas) which the research results consistently indicate to be desirable. It seems to me defensible to evaluate such teachers unfavorably. Also, we ought not to hesitate to evaluate teachers and other certificated personnel on behavior dimensions that reflect our values, regardless of their correlations with criteria of achievement.

---

[*] Barak Rosenshine, Teaching behaviors and student achievement. (New York: Humanities Press, 1971).

[†] Robert W. Heath and Mark A. Nielson, The myth of performance-based teacher education. (Berkeley, Calif.: Nomos Institute, Box 941, 1973), multilith.

If we want decency and courtesy in our teachers, then we should judge them on that basis. As we gain experience with the use of such observation instruments, the standards of desirable behavior will be clarified.

8. Don't regard a Stull Act evaluation as successful and worthwhile if it leads to the dismissal of a teacher, administrator, or counselor. Instead, explore the host of approaches available for the improvement of teachers and other certificated personnel.

Drs. Flanders and Morine have provided an extremely useful compendium of available teacher training materials and some suggestions for getting them used. Since Appendix B to their paper was prepared, additional compendia and catalogs have been published.* The whole field of in-service education for teachers and other certificated personnel has moved to a higher level during the last five years. The performance-based approach to teacher education has made available hundreds of training products. In varying degrees, these products can provide valid and feasible help in improving the behavior and performance of teachers and, to a lesser extent, other certificated personnel. These materials should also help us focus our evaluations on dimensions we can do something about rather than those that are not susceptible to change through training.

---

*W. Robert Houston, et al., Resources for performance-based education. (Albany, N.Y.: The University of the State of New York, the State Education Department, 1973); Florida Center for Teacher Training Materials Catalog. (Tallahasee, Fla.: Florida Department of Education, 1972); The Florida Catalog of Teacher Competencies. (Tallahassee, Fla.: Florida Department of Education, 1973);CEDaR Catalog of Selected Educational Research & Development Programs and Products, 3rd edition. (Denver, Col.: CEDaR Information Office, 1972).

9. Don't allow teachers to feel that they alone are being evaluated under the Stull Act. Instead, recognize and act upon the requirement that administrators as well as teachers must be evaluated.

The general problem of evaluating nonteaching certificated personnel, such as administrators and counselors, has been addressed by Drs. Abbott, Alkin and Baker, and Gordon, among others at the conference. In addition, the involvement of such personnel in the improvement of professional performance was discussed by Drs. Flanders and Morine. In short, several of the papers provided strong ideas for complying with these provisions of the Stull Act. Dr. Abbott has provided a thoughtful analysis of the dimensions of administrators that ought to be taken into account in evaluating their performance. Drs. Alkin and Baker broadened the conception of accountability to embrace goal and program accountability as well as the more frequently considered outcome accountability. As Dr. Gordon has emphasized, both here and in a subsequent document, "...The goals and objectives of auxiliary positions, whether administrative or special services, derive from those of direct instructional units such as classrooms."*

Thus, nonteaching certificated personnel should be evaluated in terms of the degree to which they help teachers foster the achievement of instructional objectives in the classroom. Just as the achievement of students is considered to depend in part on the performance of their teachers, so the performance of teachers must be considered to depend in part on the performance of administrators, principals and supervisors. One way to improve objectivity and fairness in such evaluations is to arrange for evaluators to judge their peers across district lines. Thus principals could be judged by principals from other school districts. In any case, teachers should be satisfied that these provisions of the Stull Act are being taken into account.

---

*Garford G. Gordon, Comprehensive systems for evaluating all certificated personnel. (Burlingame, Cal.: California Teachers Association, 1973), p. 2.

10. *Don't use student achievement of specific behavioral and subject matter objectives as the basis for the evaluation of administrators, counselors, and other nonteaching certificated personnel. Because nonteaching personnel are too remote from such an impact on the students of a given teacher, use observations and ratings of noncertificated personnel by those in a position to observe and judge them.*

The observations and ratings of administrators and counselors should be based on dimensions derived from theory and research. Such dimensions of administrator performance were described by Dr. Abbott. Similar analyses of the work and performance of principals, supervisors, counselors, and other nonteaching certificated personnel are needed. Such analyses will make possible the development of dimensions of observation and rating instruments that will make possible evaluations with some claim to validity. Materials and programs for the improvement of the performance of nonteaching personnel will also be needed as this kind of accountability gains momentum.

11. *Don't expect all identifiable shortcomings of certificated personnel to be remediable in the present state of the art of in-service education for teachers and others. Instead, recognize that inadequacies in teaching, administration, counseling, etc.--and the lower student achievement to which these inadequacies can contribute--cannot always be completely rectified as matters now stand.*

Just as medicine was not always able to cure pneumonia and, even now, lacks effective treatments for many diseases, in-service education for teachers and other certificated personnel is still inadequate in many respects. Physicians continued to treat pneumonia during all the decades that preceded the discovery of effective medicine. Similarly, educators will continue to try to overcome the shortcomings of teachers and other certificated personnel. But educators should not be penalized with compulsory dismissals, or even lesser threats, for their failure to bring all groups of students up to desirable levels of achievement. To do so would be much the same as punishing physicians now for their inability to cure cancer, schizophrenia, or diabetes. Again, the criterion must be that of whether a reasonable effort, in light of the state of the art, is being made.

12. Don't expect fully adequate compliance with the Stull Act during its first few years of existence. Instead, regard efforts to comply with the Act as necessarily evolving, improving as experience is gained, and requiring time and money before compliance is achieved at a level that will improve education.

Several authors and discussants of the papers presented at this conference commented on the magnitude of the task imposed by the Stull Act. They also referred to the fact that only $100,000 had been allocated by the State for the purpose of assisting in compliance with the Act. Most qualified judges estimate that adequate compliance with the letter and spirit of the Act would require expenditures of many times this amount. Beyond money, experience is required.

Presumably, the Act will be enforced with all reasonable patience and understanding of the difficulties of full compliance. The Stull Act has moved California into unknown territory. It may be that no one fully grasped the size of the job when the Act was passed. Even now, this conference has probably revealed only a small part of it. Impatient and unreasonable insistence upon perfect compliance will only compound the difficulties of doing justice to the purposes of the Act. Even if everything now known were fully applied, in all the contributing sub-disciplines of educational measurement and evaluation, we would fall far short of achieving those purposes. Even if fully adequate funds were made available, in the millions estimated to be needed, we could not now do what Mr. Stull intended. Only if we are given the time, money, and manpower will we be able to get appreciably closer. All of this considered opinion leads to my final recommendation.

13. Don't accept the Stull Act as perfect and unalterable, i.e., as neither needing change nor susceptible to change. Rather, campaign to have the Stull Act amended to provide for:

(a) Pilot tryouts in a small but representative sample of school districts to determine the feasibility of promising approaches and the costs of adequate compliance.

(b) A State advisory commission including measurement experts as well as representatives of teachers, administrators, and school boards, and

(c) A context emphasizing the improvement rather than the dismissal of teachers and other certificated personnel.

The experimental approach to the tryout of new social and educational programs has been adopted by the United States Congress and other governmental bodies. They have set up experiments in which, for example, new income-maintenance programs have been used as the independent variables and the motivation and employability of citizens have been measured as the dependent variables. The experiments have been scheduled over enough years to permit adequate follow-up of the persons involved so as to obtain valid estimates of the effects of different kinds of income-maintenance on what happens to adults and their children. Similar experiments have been used in evaluating changes in highways and traffic laws, preschool compensatory educational programs like Head Start and Follow Through, manpower programs, and programs for the treatment and prevention of drug addiction. Indeed, a whole sub-discipline of the behavioral sciences has grown up around the problem of evaluating social programs.*

As Dr. Forbes put it, "Why, in a state as sophisticated as California, didn't they select a few pilot schools to work with the claims of criteria that are being related to you here at this conference before they jumped in with both feet and got the whole state embroiled in this situation? If the motivation was really to improve education and to improve teacher and administrative performance, this seems a logical first step to have taken" (pp. 328-329).

Following models of the kinds developed for other educational and social programs, the State should authorize a series of experiments in which the provisions of the Stull Act would be carried out in real school districts. These experiments would yield rigorously validated information on costs and benefits. Just as drugs and aircraft are evaluated before they are produced in quantity, so ways of complying with the Stull Act ought to be evaluated experimentally before the Act is rigorously enforced.

The colleges, universities, school districts, professional and lay organizations, and the Department of Education of the State of California are rich in the technical expertise required for all aspects of this kind of experimentation. We have some of the world's leading authorities on educational measurement and evaluation, the formulation of objectives, experimental design and analysis, the measurement of achievement and attitudes, the administration of school districts, and the work of teaching, counseling, and administration. Some of those experts ought to be carefully chosen and formed into a commission that would develop the plans for the experimental evaluations. These plans should then be subjected to thorough examination and criticism by other experts, including representatives of the Department of Education and other appropriate bodies of the State Government. Then the experiments should be carried out over several years. Finally, the results should be analyzed and published. The experiments would yield findings that could guide the Legislature in improving upon the Stull Act. In the process, instruments and materials that would help the schools in doing their job would be developed.

---

*Peter H. Rossi and Walter Williams (eds.), Evaluating social programs: Theory, practice, and politics. (New York: Seminar Press, 1972).

In sum, the Stull Act ought to be regarded as a beginning rather than the final word. If mistakes have been made in the past, it is not too late to rectify them. If we want an answer to the question "How well does, or could, the Stull Act work?" we ought to use well-established techniques for the evaluation of social programs to find out. These techniques entail conceptual, methodological, bureaucratic, political, and organizational problems. Evaluation of the Stull Act could benefit from the experience already acquired elsewhere in performing large-scale evaluative research. This kind of research is needed to determine what the Stull Act does to schools, teachers, and pupils. Then we can proceed intelligently to the task of improving on the Act.

APPENDIX: ASSEMBLY BILL NO. 293

CHAPTER 361

An act to amend Sections 13403, 13404, 13404.5, 13405, 13406, 13407, 13408, 13409, 13410, 13412, and 13439 of, to add Sections 13413, and 13414 to, to add Article 5.5 (commencing with Section 13485) to Chapter 2 of Division 10 of, and to repeal Sections 13413, 13414, 13415, 13416, 13417, 13418, 13419, 13420, 13421, 13422, 13423, 13424, 13425, 13426, 13427, 13428, 13429, 13430, 13431, 13432, 13433, 13434, 13435, 13436, 13437, 13438, and 13440 of, the Education Code, relating to certificated employees.

[Approved by Governor July 20, 1971. Filed with Secretary of State July 20, 1971.]

The people of the State of California do enact as follows:

SECTION 1. Section 13403 of the Education Code is amended to read:

13403. No permanent employee shall be dismissed except for one or more of the following causes:

(a) Immoral or unprofessional conduct.

(b) Commission, aiding, or advocating the commission of acts of criminal syndicalism, as prohibited by Chapter 188, Statutes of 1919, or in any amendment thereof.

(c) Dishonesty.

(d) Incompetency.

(e) Evident unfitness for service.

(f) Physical or mental condition unfitting him to instruct or associate with children.

(g) Persistent violation of or refusal to obey the school laws of the state or reasonable regulations prescribed for the government of the public schools by the State Board of Education or by the governing board of the school district employing him.

(h) Conviction of a felony or of any crime involving moral turpitude.

(i) Violation of Section 9031 of this code or conduct specified in Section 1028 of the Government Code, added by Chapter 1418 of the Statutes of 1947.

(j) Violation of any provision in Sections 12952 to 12958, inclusive, of this code.

(k) Knowing membership by the employee in the Communist Party.

SEC. 2. Section 13404 of the Education Code is amended to read:

Source: Statutes of California, 1971. Pp. 720-27.

13404. Upon the filing of written charges, duly signed and verified by the person filing them, with the governing board of the school district, or upon a written statement of charges formulated by the governing board, charging that there exists cause for the dismissal of a permanent employee of the district, the governing board may, upon majority vote, except as provided in this article if it deems the action necessary, give notice to the permanent employee of its intention to dismiss him at the expiration of 30 days from the date of service of the notice, unless the employee demands a hearing as provided in this article.

Any written statement of charges of unprofessional conduct or incompetency shall specify instances of behavior and the acts or omissions constituting the charge so that the teacher will be able to prepare his defense. It shall, where applicable, state the statutes and rules which the teacher is alleged to have violated, but it shall also set forth the facts relevant to each occasion of alleged unprofessional conduct or incompetency.

Sec. 3. Section 13404.5 of the Education Code is amended to read:

13404.5. No report on the fitness of a teacher in a dismissal proceeding shall be received from a statewide professional organization by a governing board unless the teacher shall have been given, prior to the preparation of the report in its final form, the opportunity to submit in writing his or her comments on the report and unless a copy of the report in final form is given to the teacher investigated at least 10 days prior to its submission to the board.

Such a report shall not be distributed other than to the governing board and those persons participating in its preparation, unless the teacher does not demand a hearing as provided by Section 13406.

Sec. 4. Section 13405 of the Education Code is amended to read:

13405. The notice shall not be given between May 15th and September 15th in any year. It shall be in writing and be served upon the employee personally or by United States registered mail addressed to him at his last known address. A copy of the charges filed, together with a copy of the provisions of this article, shall be attached to the notice.

Sec. 5. Section 13406 of the Education Code is amended to read:

13406. If the employee does not demand a hearing by filing a written request for hearing with the governing board, he may be dismissed at the expiration of the 30-day period.

Sec. 6. Section 13407 of the Education Code is amended to read:

13407. The governing board of any school district shall not act upon any charges of unprofessional conduct or incompetency unless during the preceding term or half school year prior to the date of the filing of the charge, and at least 90 days prior to the date of the filing, the board or its authorized representative has given the employee against whom the charge is filed, written notice of the unprofessional conduct or incompetency, specifying the nature thereof with such specific instances of behavior and with such particularity as to furnish the employee an opportunity to correct his faults and overcome the grounds for such charge. The written notice shall include the evaluation made pursuant to Article 5.5 (commencing with Section 13485) of this chapter. "Unprofessional conduct" and "incompetency" as used in this section means, and refers only to, the unprofessional conduct and incompetency particularly specified as a cause for dismissal in Section 13403 and does not include any other cause for dismissal specified in that section.

SEC. 7. Section 13408 of the Education Code is amended to read:

13408. Upon the filing of written charges, duly signed and verified by the person filing them with the governing board of a school district, or upon a written statement of charges formulated by the governing board, charging a permanent employee of the district with immoral conduct, conviction of a felony or of any crime involving moral turpitude, with incompetency due to mental disability, with willful refusal to perform regular assignments without reasonable cause, as prescribed by reasonable rules and regulations of the employing school district, with violation of Section 9031, with knowing membership by the employee in the Communist Party or with violation of any provision in Sections 12952 to 12958, inclusive, the governing board may, if it deems such action necessary, immediately suspend the employee from his duties and give notice to him of his suspension, and that 30 days after service of the notice, he will be dismissed, unless he demands a hearing.

If the permanent employee is suspended upon charges of knowing membership by the employee in the Communist Party or for any violation of Section 9031, 12952, 12953, 12954, 12957, or 12958, he may within 10 days after service upon him of notice of such suspension file with the governing board a verified denial, in writing, of the charges. In such event the permanent employee who demands a hearing within the 30-day period shall continue to be paid his regular salary during the period of suspension and until the entry of the decision of the Commission on Professional Competence, if and during such time as he furnishes to the school district a suitable bond, or other security acceptable to the governing board, as a guarantee that the employee will repay to the school district the amount of salary so paid to him during the period of suspension in case the decision of the Commission on Professional Competence is that he shall be dismissed. If it is determined that the employee may not be dismissed, the school district shall reimburse the employee for the cost of the bond.

SEC. 8. Section 13409 of the Education Code is amended to read:

13409. Whenever any certificated employee of a school district is charged with the commission of any sex offense as defined in Section 12912 by complaint, information or indictment filed in a court of competent jurisdiction, the governing board of the school district shall immediately place the employee upon compulsory leave of absence for a period of time extending for not more than 10 days after the date of the entry of the judgment in the proceedings. The governing board of the school district may extend the compulsory leave of absence of the employee beyond such period by giving notice to the employee within 10 days after the entry of judgment in the proceedings that the employee will be dismissed at the expiration of 30 days from the date of service of the notice, unless the employee demands a hearing as provided in this article.

Any employee placed upon compulsory leave of absence pursuant to this section shall continue to be paid his regular salary during the period of his compulsory leave of absence if and during such time as he furnishes to the school district a suitable bond, or other security acceptable to the governing board, as a guarantee that the employee will repay to the school district the amount of salary so paid to him during the period of the compulsory leave of absence in case the employee is convicted of such charges, or fails or refuses to return to service following an acquittal of the offense or dismissal of the charges. If the employee is acquitted of the offense, or the charges against him are dismissed, the school district shall reimburse the employee for the cost of the bond upon his return to service in the school district.

If the employee does not elect to furnish bond, or other security acceptable to the governing board of the district, and if the employee is acquitted of the offense, or the charges against him are dismissed, the school district shall pay to the employee his full compensation for the period of the compulsory leave of absence upon his return to service in the school district.

Whenever any certificated employee of a school district is charged with the commission of any narcotics offense as defined in Section 12912.5, or a violation of subdivision 1 of Section 261 of the Penal Code, Sections 11530 to 11532, inclusive, 11540, or 11910 to 11915, inclusive, insofar as such sections relate to subdivision (c) of Section 11901, of the Health and Safety Code, by complaint, information, or indictment filed in a court of competent jurisdiction, the governing board of the school district may immediately place the employee upon compulsory leave in accordance with the procedure in this section.

SEC. 9. Section 13410 of the Education Code is amended to read:

13410. The notice of suspension and intention to dismiss, shall be in writing and be served upon the employee personally or by United States registered mail addressed to the employee at his last known address. A copy of the charges filed, together with a copy of the provisions of this article, shall be attached to the notice. If the employee does not demand a hearing within the 30-day period, he may be dismissed upon the expiration of 30 days after service of the notice.

SEC. 10. Section 13412 of the Education Code is amended to read:

13412. When any employee who has been served with notice of the governing board's intention to dismiss him demands a hearing, the governing board shall have the option either (a) to rescind its action, or (b) schedule a hearing on the matter.

SEC. 11. Section 13413 of the Education Code is repealed.

SEC. 12. Section 13413 is added to the Education Code, to read:

13413. In the event a hearing is requested by the employee, the hearing shall be commenced within 60 days from the date of the employee's demand for a hearing. The hearing shall be conducted and a decision made in accordance with Chapter 5 (commencing with Section 11500) of Part 1 of Division 3 of Title 2 of the Government Code, and the Commission on Professional Competence shall have all the power granted to an agency therein.

No witness shall be permitted to testify at the hearing except upon oath or affirmation. No testimony shall be given or evidence introduced relating to matters which occurred more than four years prior to the date of the filing of the notice. Evidence of records regularly kept by the governing board concerning the employee may be introduced, but no decision relating to the dismissal or suspension of any employee shall be made based on charges or evidence of any nature relating to matters occurring more than four years prior to the filing of the notice.

In those causes specified in subdivisions (b), (f), (h), (i), (j), and (k) of Section 13403, the hearing shall be conducted by a hearing officer whose decision shall be binding on the board. In the event the employee is charged with any of the causes specified in subdivisions (a), (c), (d), (e), and (g) of

Section 13403, the hearing shall be conducted by a Commission on Professional Competence. One member of the panel shall be selected by the employee, one member shall be selected by the governing board, and one member shall be a hearing officer of the State Office of Administrative Procedure who shall be chairman and a voting member of the competency panel and shall be responsible for assuring that the legal rights of the employee are protected at the hearing. If either the governing board or the employee for any reason fails to select a commission member at least seven days prior to the date of the hearing, such failure shall constitute a waiver of the right to selection, and the county board of education or its specific designee shall immediately make the selection. When the county board of education is also the governing board of the school district, the selection shall be made by the Superintendent of Public Instruction, who shall be reimbursed by the school district for all costs incident to the selection.

The member selected by the governing board and the member selected by the employee shall have at least five years' experience in the specific educational function of the accused as set forth in Section 13055.

In those instances where the employee has been charged with any of the causes specified in subdivisions (a), (c), (d), (e), and (g) of Section 13403, the decision shall be made, by a majority vote, by the Commission on Professional Competence which shall prepare a written decision containing findings of fact, determinations of issues and a disposition either:

(a) That the employee should be dismissed.

(b) That the employee should not be dismissed.

In those instances where the employee has been charged with any of the causes specified in subdivisions (a), (c), (d), (e), and (g) of Section 13403, the decision of the Commission on Professional Competence shall be deemed to be the final decision of the governing board.

The board may adopt from time to time such rules and procedures not inconsistent with provisions of this section, as may be necessary to effectuate this section.

The governing board and the employee shall have the right to be represented by counsel.

If the governing board orders the dismissal of the employee, the governing board and the employee shall share equally the expenses of the hearing, including the cost of the hearing officer. The employee and the governing board shall pay their own attorney fees.

If the governing board orders that the employee not be dismissed, the governing board shall pay all expenses of the hearing, including the cost of the hearing officer, and reasonable attorney fees incurred by the employee.

Sec. 13. Section 13414 of the Education Code is repealed.

Sec. 13.5. Section 13414 is added to the Education Code, to read:

13414. The decision of the Commission on Professional Competence may, on petition of either the governing board or the employee, be reviewed by a court of competent jurisdiction in the same manner as a decision made by a hearing officer under Chapter 5 (commencing with Section 11500) of Part 1 of Division 3 of Title 2 of the Government Code. The court, on review, shall exercise its independent judgment on the evidence. The proceeding shall be set for hearing at the earliest possible date and shall take precedence over all other cases, except older matters of the same character and matters to which special precedence is given by law.

Sec. 14. Section 13415 of the Education Code is repealed.

Sec. 15. Section 13416 of the Education Code is repealed.

Sec. 16. Section 13417 of the Education Code is repealed.

Sec. 17. Section 13418 of the Education Code is repealed.

Sec. 18. Section 13419 of the Education Code is repealed.

Sec. 19. Section 13420 of the Education Code is repealed.

Sec. 20. Section 13421 of the Education Code is repealed.

Sec. 21. Section 13422 of the Education Code is repealed.

Sec. 22. Section 13423 of the Education Code is repealed.

Sec. 23. Section 13424 of the Education Code is repealed.

Sec. 24. Section 13425 of the Education Code is repealed.

Sec. 25. Section 13426 of the Education Code is repealed.

Sec. 26. Section 13427 of the Education Code is repealed.

Sec. 27. Section 13428 of the Education Code is repealed.

Sec. 28. Section 13429 of the Education Code is repealed.

Sec. 29. Section 13430 of the Education Code is repealed.

Sec. 30. Section 13431 of the Education Code is repealed.

Sec. 31. Section 13432 of the Education Code is repealed.

Sec. 32. Section 13433 of the Education Code is repealed.

Sec. 33. Section 13434 of the Education Code is repealed.

Sec. 34. Section 13435 of the Education Code is repealed.

Sec. 35. Section 13436 of the Education Code is repealed.

Sec. 36. Section 13437 of the Education Code is repealed.

Sec. 37. Section 13438 of the Education Code is repealed.

Sec. 38. Section 13439 of the Education Code is amended to read:

13439. If the employee has been suspended pending the hearing, he shall be reinstated within five days after the governing board's decision in his favor, and shall be paid full salary by the governing board for the period of his suspension.

Sec. 39. Section 13440 of the Education Code is repealed.

Sec. 40. Article 5.5 (commencing with Section 13485) is added to Chapter 2 of Division 10 of the Education Code, to read:

Article 5.5. Evaluation and Assessment of Performance of Certificated Employees

13485. It is the intent of the Legislature to establish a uniform system of evaluation and assessment of the performance of certificated personnel within each school district of the state. The system shall involve the development and adoption by each school district of objective evaluation and assessment guidelines.

13486. In the development and adoption of these guidelines and procedures, the governing board shall avail itself of the advice of the certificated instructional personnel in the district's organization of certificated personnel.

13487. The governing board of each school district shall develop and adopt specific evaluation and assessment guidelines which shall include but shall not necessarily be limited in content to the following elements:

(a) The establishment of standards of expected student progress in each area of study and of techniques for the assessment of that progress.

(b) Assessment of certificated personnel competence as it relates to the established standards.

(c) Assessment of other duties normally required to be performed by certificated employees as an adjunct to their regular assignments.

(d) The establishment of procedures and techniques for ascertaining that the certificated employee is maintaining proper control and is preserving a suitable learning environment.

13488. Evaluation and assessment made pursuant to this article shall be reduced to writing and a copy thereof shall be transmitted to the certificated employee not later than 60 days before the end of each school year in which the evaluation takes place. The certificated employee shall have the right to initiate a written reaction or response to the evaluation. Such response shall become a permanent attachment to the employee's personnel file. Before the end of the school year, a meeting shall be held between the certificated personnel and the evaluator to discuss the evaluation.

13489. Evaluation and assessment of the performance of each certificated employee shall be made on a continuing basis, at least once each school year for probationary personnel, and at least every other year for personnel with permanent status. The evaluation shall include recommendations, if necessary, as to areas of improvement in the performance of the employee. In the event an employee is not performing his duties in a satisfactory manner according to the standards prescribed by the governing board, the employing authority shall notify the employee in writing of such fact and describe such unsatisfactory performance. The employing authority shall thereafter confer with the employee making specific recommendations as to areas of improvement in the employee's performance and endeavor to assist him in such performance.

SEC. 41. Sections 1 to 39 of this act shall become operative on the 61st day after the final adjournment of the 1972 Regular Session of the Legislature.

SEC. 42. Article 5 (commencing with Section 13401) and Article 5.5 (commencing with Section 13485) of Chapter 2 of Division 10 of the Education Code shall not apply to certificated employees in community colleges if Senate Bill No. 696 or Assembly Bill No. 3032 is enacted at the 1971 Regular Session of the Legislature.